Evaluations of Thermal Imaging Technology for Automotive Use Cases

NUI Galway
OÉ Gaillimh

Muhammad Ali Farooq (19234011)

This dissertation is submitted in fulfillment of the requirement for the degree of *Doctor of Philosophy. (Electrical and Electronic Engineering).*

Supervisor: Professor Dr. Peter Corocran

April 2022

“We crave for new sensations but soon become indifferent to them. The wonders of yesterday are today common occurrences”

~NIKOLA TESLA – My Inventions in Electrical Experimenter (1919)

Table of Contents

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Muhammad Ali Farooq
April 2022

Acknowledgments

To adequately express my gratitude and acknowledge everyone who deserves to be mentioned here would take me to go beyond the word limit for this thesis, so I want to thank everyone that helped me, in some way or another, to reach the position of submitting this Ph.D. thesis. It represents the culmination of many years of hard work.

First, I would like to thank ALLAH (SWT) who has given me enough strength to undertake this work.

A very special thanks go to my supervisor, Prof. Peter Corcoran for his guidance, helpful support, valuable and timely feedback, and suggestions to further polish my draft publications. Moreover, I would like to thank you for getting me all the hardware stuff, GPU laptop, and helping me out in collecting out-cabin thermal data, and giving me a wonderful ride in his electric car.

I would like to thank my industrial mentor, Prof. Christopher Dainty, from Xperi-Ireland. It is rare to have the opportunity to interact with a scholar of your accomplishments and yet you always had time for me.

I would like to thank my French, Italian, and German industrial partners who were participating as consortium partners in Heliaus project especially: Quentin Noir, Guillaume Delubac, Merla Arcangelo, Lorenza, and Benedikt Prähofer for sharing their useful feedback and doing impactful collaboration throughout my Ph.D.

Thank you to all my Xperi colleagues especially: Dr. Amr Elrasad, Dr. Cain Ryan, Dr. Joseph, Lemley, David O'Callaghan, Cosmin Rotaria for helping me out with project-related experimental and data acquisition work.

A special thank goes to my besties Dr. Faisal Khan and Dr. Waseem Shariff for all the wonderful moments we spent together, helping me out in paper drafting and proofreading the work, all the lunches and dinners we did together, funny, and sometimes useful conversations, and jokes we crack on each other.

I would like to thank all my Ph.D. group friends Dr. Adrian Stefan Ungureanu, Dr. Asma Khatoon, Dr. Yahya Khan, Dr. Shubhajit Basak, Dr. Dan Bigioi, Dr. Wang Yao, Dr. Rishab Jain, and Dr. Mehdi for all the wonderful times we spent together. I will miss out group very much!

To Dr. Claudia Costache and Dr. Marium Yaware, for being the best postdocs in the world. For always being there with a kind word and helpful advice, technical or otherwise.

Thank you to Joe Desbonnet for all the support with hardware stuff!

Thank you to all the employees of Fotonation (Galway offices), for the constructive conversations and all the collaborations throughout these years.

I want to thank the entire group in the Electronic Engineering department in NUIG, especially: Dr. Evismar Andrade, Dr. Hao Lin, Meehan Myles, and Martin Burke.

A warm thank you to my postgraduate society group members with whom I shared countless happy hours and got familiar with people from diversified cultural groups.

A very special thanks to all my previous and hardworking faculty members and teachers especially my Master's supervisor Dr. Hammad Raza for their hard work and for helping me out throughout my academic career and motivating me to become an independent researcher.

Last but not least the most special and heart-warming acknowledgment and thanks go to my loving and caring family members: my mother Munira Bano and my father Farooq Ibrahim Naviwala, my brothers Murtaza Farooq, my grandfather Ibrahim Hameed Naviwala, and all other family members for all their support, and countless prayers for my success.

Abstract

Thermal imaging has been widely used in high-end applications for instance industrial and military applications as it provides superior and effective results in challenging environments and weather conditions such that in low lighting scenarios and has aggregate immunity to visual limitations thus providing increased situational awareness. This research is about exploring the potential of thermal imaging for smart vehicular systems including both in-cabin and out-cabin applications using uncooled LWIR thermal imaging technology. Novel thermal datasets are collected in indoor and road-side environments using an especially designed low-cost, yet effective prototype thermal camera module developed under the Heliaus project.

The collected data along with public datasets are further used for generating large-scale thermal synthetic data using the composite structure of advanced machine learning algorithms. The next phase of this work focuses on designing AI-based smart imaging pipelines which include driver gender classification system and object detection in the thermal spectrum. The performance of these systems is evaluated using various quantitative metrics which include overall accuracy, sensitivity, specificity, precision, recall curve, mean average precision, and frames per second.

Furthermore, the trained and fine-tuned neural architectures on thermal data are deployed on Edge-GPU embedded devices for real-time onboard feasibility validation tests. This is accomplished by performing optimal optimization of successfully converged deep learning models on thermal data using SoA neural accelerators to achieve a reduced amount of inference time and a higher FPS rate.

List of Figures

List of Tables

Abbreviations

AI: Artificial Intelligence

LWIR: Long Wave InfraRed

ANN: Artificial Neural Network

CNN: Convolution Neural Network

DNN: Deep Neural Network

SVM: Support Vector Machines

EU: European Union

DMS: Driver Monitoring System

ADAS: Advanced Driver Assistance System

CMOS: Complementary Metal Oxide Semiconductor

CPU: Central Processing Unit

GPU: Graphical Processing Unit

TPU: Tensor Processing Unit

FPGA: Field Programmable Gate Arrays

ASIC: Application-Specific Integrated Circuits

YOLO: You Only Look Once

IRT: Infrared Thermography

RGB: Red Green Blue

FLIR: Forward Looking InfraRed

FPS: Frames Per Second

AGC: Automatic Gain Control

BPR: Bad Pixel Removal

TD: Temporial Denoiseing

VGA: Video Graphics Array

2D: Two Dimensional

3D: Three Dimensional

GAN: Generative Adversarial Network

CT: Computed Tomograpy

Chapter 1

Introduction

Thermal infrared imaging systems based on microbolometer technology are the common sensor of choice for many imaging applications, including the automotive sensor suite for designing smart in-cabin and outdoor thermal perception systems. A microbolometer is an array of tiny heat-detecting sensors attached to a special type of lens that is sensitive to infrared radiations and can measure infrared thermal emissions emanating from the object thus such types of cameras don't rely on reflected light. They can provide high-resolution imagery in the day as well as nighttime even in zero lighting conditions. This research work is about exploring the potential use of uncooled thermal cameras based on microbolometer technology for in-cabin and out-cabin automotive applications thus enabling safe driving systems.

This Ph.D. research work is a part of and correlated with the Heliaus EU H2020 project [1]. Heliaus project is about exploring the potential usage of low cost yet effective uncooled thermal imaging modules for smart automotive applications. Taking the advantage of Long Wave Infra-Red bandwidth (LWIR) in which all bodies emits energy depending on their structure and temperature, the core focus of this project is to design and deliver breakthrough thermal perception systems. Such types of advanced photonics-based systems will be effective and beneficial for both in-cabin passenger monitoring and for observing the car surroundings thus enabling enhanced safety features for automotive applications. Figure 1 shows the general description (block diagram) representation of the Heliaus project. The manufacturing of thermal imaging cameras undergoes a systematic, yet complicated, production process to integrate the parts and it offers a wide range of benefits over conventional CMOS imaging devices. The operational capabilities of thermal sensors are not affected by wind, moisture, rain, and other such types of harsh weather conditions. As compared to visible imaging cameras that depend on external lighting conditions to capture images, the thermal imaging module absorbs heat radiated from the body, allowing them to detect even distant objects in both day and nighttime conditions. Moreover, thermal cameras are capable of producing clear, accurate images, resulting in fewer false alarms. In this industry research project [1] we were involved in different work packages of this project which include work package 7 (WP-7), work package 8 (WP-8), and work package 4 (WP-4). As a part of these work packages, we worked on various tasks which includes in-door and out-cabin thermal data collection, synthetic thermal data generated using the composite structure of computer vision algorithms, face localization in thermal images, autonomous driver gender classification, and thermal object detection framework. The main focus of the out-cabin dataset acquisition is to further use that dataset to develop efficient object detection and classification framework for the automotive sensor suite that can work precisely in all light conditions, provide redundancy, and thus extends vehicle autonomy. As a consortium member of this industry-based Ph.D. program, we worked closely with different industrial partners which include Xperi, Lynred, Next2U, and Denso who were all involved and working as consortium partners in this EU-funded Project in solving real-world problems for automative technologies using advanced machine learning algorithms.

The use of thermal imaging and further feeding that data for training machine learning algorithms to build AI-based solutions can lead towards developing effective smart thermal perceptions systems as shown in Figure 1 for advanced vehicular systems. AI algorithms especially deep learning methods have gained much popularity nowadays due to their robust outputs for various real-world applications. Deep learning is an emerging area in the field of machine learning and has been introduced to achieve the goals of enhanced and improved prediction accuracy levels as compared to conventional machine learning methods. Deep learning architectures or deep neural networks (DNN) are the further extension of artificial neural networks and play a vital role to design artificially intelligent imaging pipelines. Normally the neural classifiers use one or two hidden layers of neurons, and they are most used for supervised machine learning tasks but on the other hand, deep neural architectures consist of a greater number of hidden layers which can range from three hidden layers to up till several hidden layers. Deep neural networks differ from conventional machine learning algorithms such as Support Vector Machines (SVM) as they can be trained using both supervised and unsupervised learning algorithms. Deep learning methodologies are used in a wide range of applications which includes digital signal processing, image classification, and sound classification. Due to their precise and robust results, these types of frameworks are widely used in complex image understanding and processing applications which include satellite imaging, medical image analysis, human biometrics, and other such types of tasks.

In this research work, we have particularly focused on supervised learning methodology and used different types of CNN architectures for various in-cabin and out-cabin applications which include human thermography, designing thermal gender classification system, thermal object detection/ classification framework, and synthetic data generation. Moreover, the further stage of this research work focuses on the deployment of trained/ fine-tuned networks on single board edge-GPU devices for onboard real-time feasibility testings.

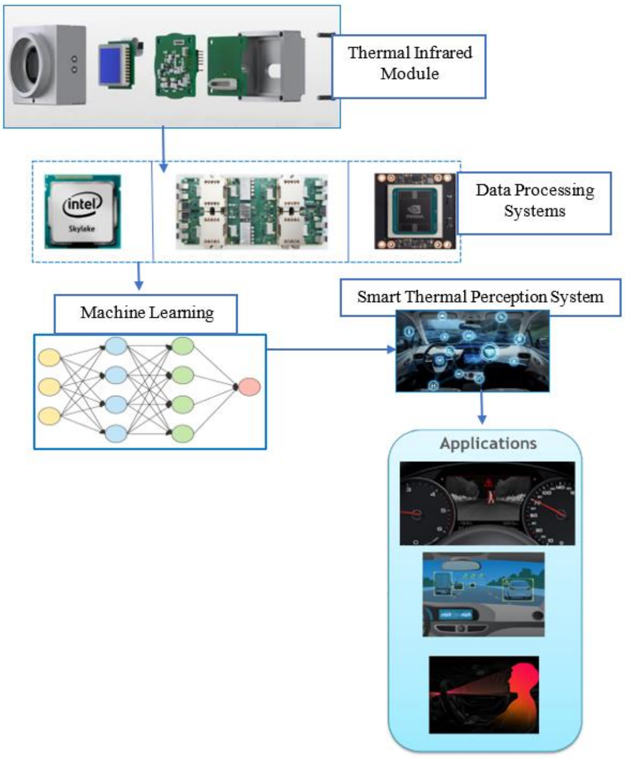


Figure 1: General description (block diagram representation) of Helius Project

1.1. Objectives and Scope of the Work

With the rise of new, often disruptive forms of personal mobility, the role of the automotive industry, and its engagement with stakeholders and end-users, as a whole is changing rapidly. Drivers will demand a better driving experience for themselves, their passengers, and other road users. The advancement in photonic technologies plays a critical role in meeting these requirements and expectations. The main objective of the Heliaus project is to explore the feasibility of improved technologies and systems in the field of thermal sensing technologies. These technologies will be advantageously aggregated into functional small-scale prototypes. These prototypes will then be used to establish augmented perception systems for a wide range of transport and smart mobility applications. This thesis is drafted in the context of the Heliaus project and discusses the role of thermal infrared imaging by integrating it with advanced machine learning algorithms for the development and deployment of effective thermal perception systems as a viable solution for drive monitoring systems (DMS) and advanced driver-assistance systems (ADAS).

1.1.1 Intelligent Vehicular Systems

The new ways of driving and using a vehicle as expected in the scope of smart mobility ask for reliable, affordable, and versatile perception systems. It is now clear that the interaction between the vehicles and environment (in and out cabin) has to be improved. Over the past few years, various safety features have been deployed into cars, including adaptive cruise control, driver monitoring systems, lane departure warning, blind-spot detection, out-cabin object detection, tracking, and other such intelligent functions. These systems play an important role to increase road safety by helping the drivers in smart manner [2]. These intelligent systems work by taking real-time data information from various optical as well as typical hardware sensors such as Lidar and Radar [3]. In this work, we have specifically focused on using low-cost, yet effective uncooled thermal cameras based on microbolometer sensors as a heat detector for various computer vision applications that can be integrated with other intelligent systems for the automotive sensor suite.

1.1.2 Smart Thermal Perception Systems

The Heliaus project integrates high-performance and low-cost thermal systems, images/ data processing units, and advanced machine learning methods into smart thermal perception systems. Such type of perception systems can be used solely or in combination with other sensors commonly used for vehicular applications, for in-cabin monitoring, or for out-of-cabin applications, leading to an augmented awareness and reliability. The recent surge of interest in machine learning especially deep learning is due to the immense popularity and effectiveness of convnets [4]. In [5] authors have done an in-depth comparison of conventional machine learning vs deep learning algorithms thus showing the effectiveness of deep learning methods. Convolution Neural Networks (CNN) architectures are commonly used for designing AI-based intelligent imaging pipelines for numerous consumer technology applications [6-7]. In this research work, we have explored and employed deep learning algorithms for thermal imaging based smart perception system for all weather and environmental conditions. The data was acquired from an uncooled LWIR thermal camera developed under the Heliaus project [1]. This can be eventually beneficial for designing intelligent imaging pipelines effective for advanced vehicular systems such as image classification, object detection, and object tracking.

1.1.3 Deployment of AI-based Thermal Imaging Pipelines on Edge Devices for Onboard Automotive Installations

The Heliaus project contributes to the development of AI solutions at the edge, as augmented perception systems for autonomous driving are targeted. Edge computing on embedded AI platforms plays a significant role in deploying and evaluating the performance of trained machine learning algorithms for real-time applications. Gartner predicts by the year 2025, edge computing will process about 75% of data generated by all use cases, including those in factories, healthcare, and transportation [8]. Embedded devices use different types of processing cores such as central processing units (CPU), graphical processing units (GPU), tensor processing units (TPU), field programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs). This study evaluates the real-time feasibility analysis of SoA thermal image inference networks by deploying a forward sensing thermal object detection system on embedded-GPU devices which includes Nvidia Jetson Nano and Nvidia Jetson Xavier for automotive applications.

1.2. Summary of The Main Contributions in This Thesis

The sections present the core contributions of this thesis which are summarized in the below sub-sections. In the remaining chapters of this thesis, the work related to these contributions is presented. In each chapter, an introductory paragraph provides the context of the research work. Following that, the research objectives of the work are given, followed by the contributions of the presented research work. The first section will explicitly summarize the WP-7 contributions which are mainly related to in-cabin applications. Whereas the next section will list our WP-8 contributions which are related to out-cabin applications. The last section will list WP-4 contributions which are related to the deployment of AI-based smart imaging pipelines on resource constraint embedded boards for on-board feasibility testings.

1.2.1 Contribution To Work Package 7 (In-cabin Applications) of Heliaus Project

The in-cabin applications development targeted in the context of the Heliaus project aim at prototyping new smart thermal systems enabling the monitoring of driver activities by specifying the person's soft biometrics, vital sign monitoring, and drowsiness detection. The main contributions related to this work package are listed below.

1. *Contribution To Indoor Thermal Data Acquisition using Protytpe LWIR Thermal Camera Module and Synthetic Thermal Data Generation using Computer Vision Methods*

This first phase contributes toward a novel indoor thermal face data which is acquired using a 640x480 prototype uncooled LWIR thermal camera. This camera module is developed under the Heliaus project [1]. The main goal of collecting this data is to further use it for various in-cabin applications which include synthetic thermal data generation for robust training of deep learning models and, the development of autonomous driver gender classification. In addition to collecting our thermal datasets as discussed in section 3.3.1 of chapter 3, I also contributed by helping our industry partner Xperi in the data collection process and participating in in-cabin data collection as a volunteer subject. The goal of acquiring this data is to observe natural drowsiness behavior and high cognitive load in a simulated driving situation with several optical and electrical sensing modalities. Furthermore, the second phase of this work includes synthetic thermal data generation using existing thermal public and locally acquired datasets by employing various computer vision algorithms. The complete working methodology and experimental details are presented in our published conference papers [9-10].

2. *Contribution to Development of In-Cabin Thermal gender Classification System using SoA CNNs*

The second important contribution is the development of an in-cabin driver gender classification system using SoA pre-trained deep learning architectures. In this experimental work [11] we have first trained the nine deep learning architectures on a large-scale CASIA facial dataset. In the second phase, the trained DNNs are further fine-tuned on Tufts thermal dataset [12-14]. The efficacy of all the thermally tuned networks is validated on unseen test data collected from the Carls thermal dataset [15-16] and locally acquired indoor thermal data. In addition to using pre-trained neural networks, a new application-specific CNN architecture, GENNet, is designed and its performance is evaluated against the nine pre-trained CNN networks. The further details of this work are summarized in chapter 4.

1.2.2 Contribution To Work Package 8 (Out-cabin Applications) of Heliaus Project

The out of cabin applications development targeted in the context of the Heliaus project aim at prototyping new smart thermal systems enabling the detection, labeling, and classification of the objects, including human beings and other road-side objects, surrounding the vehicle by using computer vision algorithms based on the use of thermal images. Moreover, the main focus is the improvement of such perception systems in the most challenging environmental or light conditions. The core contributions related to this work package are listed below.

1. *Contribution To Novel Object Detection Thermal Data Acquisition using Protytpe LWIR Thermal Camera Module*

This first phase contributes toward a novel roadside thermal object detection data which is acquired using a prototype 640x480 uncooled LWIR thermal camera [17]. The main goal of collecting this data is to further use it for out of cabin applications which includes the development of the SoA thermal object detection framework that should be effective in all weather and environmental conditions. The further details of this data are presented in section 3.3.2 of chapter 3.

2. *Contribution to Development of Out-cabin Thermal Object Detection/ Classification System using SoA YOLO Framework*

The second contribution towards this work package includes the adaptation and validation of a state-of-the-art object detection/ classification YOLO framework for designing an out-cabin smart thermal perception system with seven distinct classes including stationary as well as moving objects [18]. This includes the preparation and annotation of a large-scale locally acquired dataset of thermal images captured in different weather and environmental conditions for out-cabin object detection. Moreover, a new model ensemble-based inference engine is proposed using the combination of two best-trained models to further improve the accuracy metrics on test data. The further details of this work are summarized in chapter 5.

1.2.3 Contribution To Work Package 4 of Heliaus Project

This work package involves the efficient real-time implementation and deployment of Neural Network-based processing frameworks on dedicated embedded devices. The contribution related to this work package is listed below.

1. *<u>Contribution to Deployment of Object Detection/ Classification System on GPU & EDGE-GPU devices using Advanced Neural Optimization Methods</u>*

This major contribution to this work package incorporates further evaluating the neural framework with a range of model sizes to determine its suitability for porting to a resource-constrained embedded edge platform which includes Nvidia Jetson Nano and Nvidia Jetson Xavier embedded boards. Thus, to study its feasibility in the form of inference time required and fps rate for further automotive on-board-computer (OBC) installations. In this work [19] we have performed model optimization using the SoA TensorRT inference accelerator to implement a fast inference network on SoA embedded GPU boards (Jetson, Xavier) with comparative evaluations. The further details of this work are summarized in chapter 6.

1.2.4 Additional Contributions

The other contributions of this thesis incorporate a detailed study about the effective use of thermal imaging for human thermography [20]. The overall study emphasized the significance of Infrared Thermography (IRT) and the role of machine learning in thermal medical image analysis for human health monitoring and various disease diagnosis in preliminary stages. In the second phase, we have proposed a breast tumor classification system using thermal frames and skin cancer classification systems using RGB dermoscopic images by applying transfer learning methodology for fine-tuning the selected set of pretrained deep neural networks [20-21].

In addition to that, I have also worked on monocular depth estimation with my fellow Phd colleagues. In this work, we have explored various depth datasets, SoA algorithms for depth estimation using 2D RGB images of different scenes and environments. In addition to that, we have also benchmarked the performance of the proposed depth estimation algorithm with other DNN algorithms [22]. The further details of these additional contributions are summarized in chapter 7.

1.3. List of Publications

The work presented in this thesis resulted in the following journal and conference papers publications.

1.3.1 WP-7 Contribution Publications

This section will list the number of publications related to WP-7 of the Heliaus project. In this work package, one journal and two conference papers have been published. The copy of the published papers is attached in Appendix A, Appendix B, and Appendix C of this thesis report.

1. Farooq, Muhammad Ali, Hossein Javidnia, and Peter Corcoran. "Performance estimation of the state-of-the-art convolution neural networks for the thermal images-

based gender classification system." *Journal of Electronic Imaging* 29.6 (2020): 063004.

2. M. A. Farooq and P. Corcoran, "Generating Thermal Image Data Samples using 3D Facial Modelling Techniques and Deep Learning Methodologies," *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1-5, DOI: 10.1109/QoMEX48832.2020.9123079.

3. M. A. Farooq and P. Corcoran, "Proof-of-Concept Techniques for Generating Synthetic Thermal Facial Data for Training of Deep Learning Models," *2021 IEEE International Conference on Consumer Electronics (ICCE)*, 2021, pp. 1-6, DOI: 10.1109/ICCE50685.2021.9427690.

1.3.2 WP-8 Contribution Publications

This section will list publications related to WP-8 of the Heliaus project. In this work package, 'C3I Thermal Automotive Dataset' and one journal paper have been published. The copy of the published paper is attached in Appendix D and Appendix E of this thesis report.

4. Muhammad Ali Farooq, Waseem Shariff, Faisal Khan, Peter Corcoran, Cosmin Rotariu, March 26, 2022, "C3I Thermal Automotive Dataset", IEEE Dataport, DOI: https://dx.doi.org/10.21227/jf21-rt22.

5. M. A. Farooq, P. Corcoran, C. Rotariu and W. Shariff, "Object Detection in Thermal Spectrum for Advanced Driver-Assistance Systems (ADAS)," in *IEEE Access*, DOI: 10.1109/ACCESS.2021.3129150.

1.3.3 WP-4 Contribution Publications

This section will list publications related to WP-4 of the Heliaus project. In this work package, one journal article has been published. The copy of the published paper is attached in Appendix F of this thesis report.

6. M. A. Farooq, W. Shariff and P. Corcoran, " Evaluation of Thermal Imaging on Embedded GPU Platforms for Application in Vehicular Assistance Systems," Published in *IEEE* Transactions on Intelligent Vehicles, DOI: 10.1109/TIV.2022.3158094.

1.3.4 Additional Contribution Publications

This section will list other contribution publications which are not related to any specific work package of the Heliaus project [1], however, this work was done as additional contributions during my Ph.D. program. The copy of the published paper is attached in Appendix G, Appendix H, and Appendix I of this thesis report.

7. M. A. Farooq and P. Corcoran, "Infrared Imaging for Human Thermography and Breast Tumor Classification using Thermal Images," *2020 31st Irish Signals and Systems Conference (ISSC)*, 2020, pp. 1-6, DOI: 10.1109/ISSC49989.2020.9180164.

8. M. A. Farooq, A. Khatoon, V. Varkarakis, and P. Corcoran, “Advanced deep learning methodologies for skin cancer classification in prodromal stages,” CEUR Workshop Proc., vol. 2563, pp. 40–51, 2019.

9. F. Khan, M. A. Farooq, W. Shariff, S. Basak and P. Corcoran, "Towards Monocular Neural Facial Depth Estimation: Past, Present, and Future," in *IEEE Access*, doi: 10.1109/ACCESS.2022.3158950.

1.4. Thesis Structure

The rest of the thesis structure is as follows.

Chapter 2 introduces an overview of infrared thermal imaging, different infrared spectrums, thermal camera hardware and different types of thermal cameras, applications of thermal imaging in different sectors, and commercial manufacturers of thermal cameras. Moreover, this chapter will further provide the details about the LWIR prototype thermal camera specifically designed and used during this research work for recording indoor as well as out-cabin thermal data for various automotive applications.

Chapter 3 presents complete details about new thermal data acquisition using the prototype 640x480 uncooled thermal camera. This includes types of thermal data that are being recorded for being utilized in different types of experimental works and data collection methods with complete dataset attributes.

Chapter 4 presents our contributions towards work package 7 (WP-7) of the Heliaus project [1] which is focused on the development and validation methodologies of the thermal-IR system for in-cabin vehicular applications. In this work, we have developed an efficient thermal gender classification system using end-to-end pretrained deep learning networks as well as newly proposed GENNet convolution neural networks for in-cabin driver monitoring applications. Further details of these contributions are detailed in journal publication 1 [11] and conference publications 2 [9] & 3 [10] listed in section 1.3.1.

Chapter 5 presents our contributions towards work package 8 (WP-8) of the Heliaus project [1] which is focused on the development and validation methodologies of the thermal-IR system for out-cabin vehicular applications. In this work package, we have explored adapting and, modifying state-of-the-art object detection and classifier framework on thermal data with various distinct classes for advanced driver-assistance systems (ADAS). Further details of these contributions are detailed in journal publication 5 [18] listed in section 1.3.2.

Chapter 6 presents our contributions towards work package 4 (WP-4) of the Heliaus project which further explores object detection model optimization by presenting our core contributions towards using advanced neural methods for deploying fast inference engines on embedded architectures. Further details of these contributions are detailed in journal publication 6 [19] listed in section 1.3.3.

Chapter 7 presents additional contributions regarding human thermography and fatal disease diagnosis in the early stages using advanced deep learning algorithms. Moreover, it also highlights the research work done in collaboration with other universities and Ph.D. students.

Further details of these contributions are detailed in conference publications 7 [20], 8 [21] and journal publication 9 [22] listed in section 1.3.4.

Chapter 8 outlines the main conclusions and future work based on the work contained in this thesis.

Chapter 2

Thermal Imaging Background

This chapter will mainly focus on infrared thermal imaging, applications of thermal imaging, thermal camera hardware, and its working methodology, the difference between CMOS and thermal imaging sensors, different types of thermal cameras, commercial manufacturers of thermal cameras, and different infrared spectrums. This chapter will further introduce the LWIR prototype thermal camera specifically designed and used during this research work for recording indoor as well as out-cabin thermal data for various automotive applications. Lastly, this chapter will discuss the shutter-less camera calibration method for recording high-quality thermal images.

2.1. Thermal Infrared Imaging

Infrared radiation was initially found in 1800 by Sir Frederick William Herschel (1738-1822), who is likewise popular for finding the planet Uranus as well as composing 24 orchestras [23]. Thermal imaging is one of the most rapidly growing imaging techniques nowadays [24]. It can be described as a key method for measuring the spatial temperature of various materials, objects, and scenes. It works by absorbing IR radiations from the objects and then generating heat energy indications with or without visible illumination conditions using different colour maps such as greyscale, iron, glow, and rainbow. These color maps are generally used to define different temperature ranges which eventually help us in identifying the health parameters of different objects. Thermography is a non-invasive tool that utilizes thermal data captured from the thermal camera for acquiring useful information [25]. It collects information using an array of infrared sensors to read infrared energy emissions (surface temperature) in order to determine the operating conditions of different parts of the objects as well as the human body. It consists of two main components: Thermo and graphy where thermo refers to temperature patterns of the body and graphy refers to image acquisition techniques as shown in Figure 2. Thermal video, thermal images, and Infrared thermography are examples of infrared imaging science. Thermographic cameras usually detect radiation in the long-infrared range of the electromagnetic spectrum (roughly 9,000– 14,000 nm or 9–14 µm) and produce images using that radiation, which is generally referred to as thermograms. The amount of radiation emitted by an object increases with an increase in the temperature; therefore, thermography helps to see temperature variations in a much more visible manner.

In the general classification, thermography can be divided into two main types which include active thermography and passive thermography. Passive thermography works by pointing the IR camera at the investigated body and checking whether the investigated body is at a lower or higher temperature than the background. Whereas the active thermography approach is based on the excitation of the sample by applying external energy into it and subsequently measuring the thermal response from it. Therefore, active thermography is a fully dynamic process requiring different methods of image processing [26]. The second part of this chapter focuses on the real-world health monitoring applications of the human body using infrared thermography (IRT). The human body temperature is considered the most important vital parameter that can be used for further diagnosis of normal as well as fatal diseases.

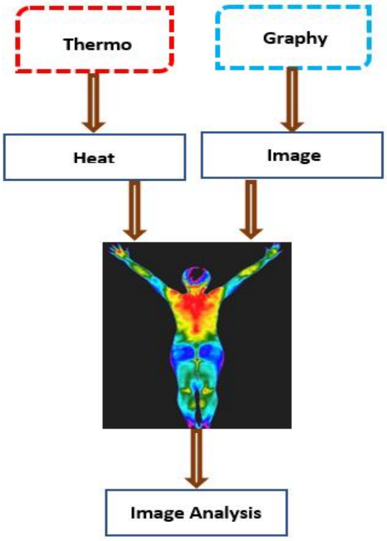


Figure 2: Main componennts of thermography

Thermography has several advantages such as it is contactless, provides high speed, is portable, durable, used for Non-Destructive Testing (NDT), and eventually, it can inspect large areas, therefore, it is applicable in a wide range of real-world applications [27]. For instance, in the health sector thermal cameras can be used to monitor different health parameters of a body such as fever measurement, and blood pressure measurement. Moreover, it can be used for disease diagnosis, cancer lesion segmentation, and infection detection [28]. In the aerospace industry, it can be used to detect Carbon/epoxy composites, delamination/impact, and monitor engine performance [29]. In Automotive Industry, it can be used for continuous monitoring of composite structures, non-destructive testing, spot welds, and adhesive bonds [30]. In the power and electrical sector, it is used for measuring the performance of wind turbine blades, coating uniformity, and delamination in composites [31]. Moreover, it can be used for an extensive range of electrical applications such as the detection of unbalanced Loads, detection of loose or corroded connections, and detection of winding insulation failure in electric motors [32]. In the defense industry, thermal cameras which are specially designed for military applications are deployed in military hardware for performing specialized operations in rough environmental conditions especially in the nighttime to achieve precision accuracy [33]. Lastly, it is also beneficial for agricultural applications as it plays a vital role in the agriculture and food industry. It is used for predicting water stress in yields, forecasting and scheduling irrigation in a good time frame, pathogen, and disease diagnosis in flowers, predicting fruit yield, evaluating the maturing of fruits, bruise detection in fruits and vegetables, and temperature distribution during cooking [34].

2.1.1 Effective Factors for Thermal Imaging

Thermal imaging is one of the most sensitive applications. Therefore, three important factors need to be considered while collecting the data or making large datasets for various real-time applications using different types of thermal cameras. These factors are described below.

1. Environmental Factors: These factors mainly depend upon outdoor and indoor environmental conditions such as ambient temperature, air, and atmospheric pressure, relative humidity, and the radiation source. Indoor environmental conditions can be controlled, however, the outdoor environment cannot be controlled due to natural ecological features.

2. Individual Factors: These factors are generally dependent on the human body's intrinsic and extrinsic characteristics that can affect the overall body temperature. Intrinsic factors are generally due to biological and anatomical constraints such as age, gender, hair density, metabolic rate, skin irradiation, genetics, blood pressure, and emotions. The extrinsic aspects that can affect body temperature include food intake, drug consumption, stimulants, cosmetics, ointment, etc. However, these factors can vary from person to person, and it is nearly impossible to control them.

3. Technical Factors: These factors are generally due to the camera electronics, electronic components noise, camera position, camera distance from the subject, calibration, selection of Region of Interest (ROI), appropriate selection of image processing and machine learning algorithm, and statistical analysis, etc. In general, as compared to environmental and individual, technical factors are controllable by using modern technological solutions.

2.2. Working Principle of Thermal Cameras

Thermal cameras are equipped with microbolometer image sensors. A microbolometer comprises thousands of tiny heat-detecting sensor elements as shown in Figure 3. Each element has a micro resistor which changes its resistance as the camera detects heat radiations. The thermal camera focuses heat onto the elements which in turn heat up. The working principle of these cameras functions by detecting the change in resistance resulting from the absorption of IR radiant energy normally within the wavelength between 9–14 um. The change in resistance is measured and then processed into temperature values which are represented graphically in the form of thermal images. The development and working methodology of different types of infrared sensors and detectors are comprehensively defined by Zhang, et al. [35]. Thermal imaging is a non-invasive imaging method acquired using thermal cameras. These cameras can measure the temperature of the body without the need for direct contact. To collect thermal images for various applications including human thermography different types of thermal cameras are used which are equipped with InfraRed (IR) or thermal detector sensors. Figure 4 displays the sample image of generic thermal cameras. Thermal cameras are available in different variants designed for specialized applications. It includes thermal cameras for drones, mobile thermal cameras for apple and android devices, scout, and thermal security cameras. These thermal detectors are generally divided into two main categories which include cooled and uncooled, respectively. The recent advancements in solid-state electronic chips, it has eventually cleared the path for the creation of more up-to-date thermal detectors with good accuracy and precision levels. Presently, the thermal affectability of the cooled cameras is about 0.05 °C contrasted with 0.01 °C of uncooled cameras.

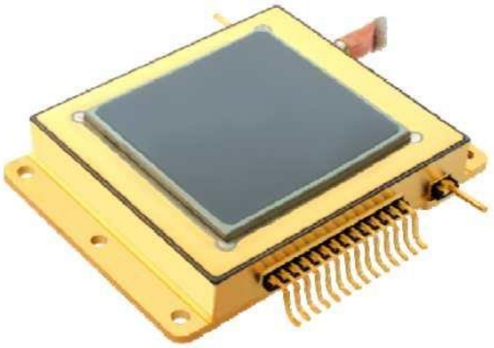


Figure 3: Microbolometer sensor array packaging [49].

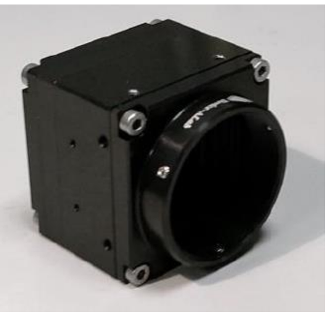


Figure 4: Sample image of the thermal camera.

These cameras have numerous favorable circumstances, including space and high-temperature accuracy and portability. Moreover, these lightweight uncooled thermal cameras are manufactured by thin-film silicon innovation, thus they are less expensive compared to cooled thermal infrared cameras. The latest uncooled thermal cameras have extensively achieved improved thermal imaging capabilities for providing exceptional performance in different real-time applications such as human thermography, machine health monitoring, high tension electrical cable inspections, and other such applications [36-38].

2.3. Difference Between CMOS Image Sensors & Thermal Image Sensors

The working principle of a Complementary Metal Oxide Semiconductor (CMOS) image sensor was established in the latter half of the 1960s. However, the device was yet not commercialized until microfabrication technologies became advanced enough in the 1990s. A Complementary Metal Oxide Semiconductor (CMOS) camera sensor as shown in Figure 5 is a type of imager that collects visible light ranging from 400~700nm band (which is the same spectrum that the human eye perceives) and converts that to an electrical signal. In the next stage, it organizes that information to render images and video streams. Image sensors assembled into today's digital/ RGB cameras and mobile phone cameras mostly use either the CCD (charge coupled device) or CMOS technology. Visible cameras are designed to create images, capturing light in red, green, and blue wavelengths (RGB) for accurate color representation. As compared to the human eye which requires visible light, RGB cameras also require light in the visible spectrum to generate images. Due to this reason the visible cameras are considered unfavorable for producing adequate outputs in low-lighting or zero lighting conditions. Their performance is also significantly reduced by harsh atmospheric conditions such as fog, haze, smoke, heat waves, and smog. This limits their usage and applications to daytime and clear weather conditions mostly.

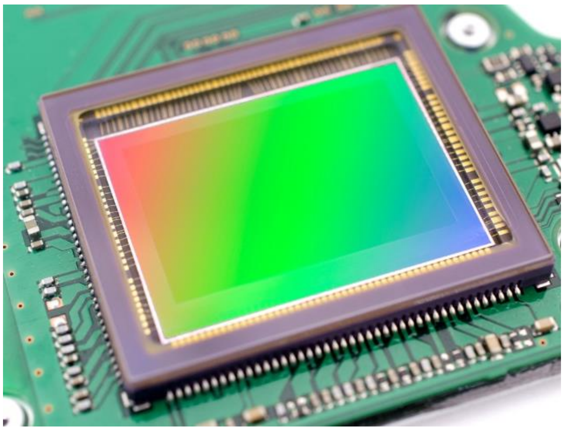


Figure 5: CMOS image sensor.

In comparison to this thermal infrared cameras do not require any additional visible light conditions to operate and have the ability to generate high contrast images even in night-time scenarios. This means that thermal cameras can be deployed very subtly while remaining highly effective. This makes thermal cameras the perfect choice for the day as well as night-time or low lux scenes applications. Moreover, the thermal cameras are highly effective in diversified environmental conditions such as fog, haze, smoke, or sandstorms that can deter the performance of visible cameras thus making them ineffective in challenging environmental conditions. In this thesis, the main reason for selecting thermal imaging technology for vehicular applications is to develop intelligent systems that should remain effective and functional irrespective of lighting conditions and provide robust results in diversified weather and environmental conditions. Table 1 shows the comparison of CMOS and thermal imaging sensing technology by analyzing various factors.

Table 1: Comparison of CMOS and Thermal IR Sensing Technology

Sensor	Thermal IR	CMOS
Pixel Signal	Electron Packet	Voltage
Chip Signal	Digital	Digital
Fill Factor	High	Moderate
Responsivity	High	Moderate-High
Noise Level	Low	Moderate – High
Dynamic Range	High	Moderate
Uniformity	High	Low
Resolution	Low – Moderate	Low – High
Speed	Moderate - High	Moderate - High
Power Consumption	Low - High	Low - Moderate
Complexity	Low	Moderate
Cost	Moderate - High	Moderate

2.3.1. Types of Thermal Cameras

Thermal cameras which are available commercially can be divided into two broad categories. It includes cooled thermal cameras and uncooled thermal cameras.

Cooled Thermal Camera: The latest cooled technology thermal cameras are equipped with an imaging sensor that is integrated with a cryocooler. It is the type of device that brings down the sensor temperature to cryogenic temperatures. It is necessary to decrease the sensor temperature to eventually reduce the thermally actuated noise to a level beneath that of the sign from the scene being imaged. Cryocoolers have moving parts made to very close mechanical tolerances that wear out after some time along with helium gas that gradually works its way past gas seals. Cooled thermal imaging technology is considered the most sensitive type of thermal imaging technique that has the ability to even detect minute temperature differences between the objects which are very useful for applications that require precise and robust results. They can generate a picture in the mid-wave infrared (MWIR) band and the long-wave infrared (LWIR) band of the range where the thermal complexity is high because of blackbody material science. The thermal difference is the adjustment in signal for an alteration in target temperature. The higher the thermal difference, the simpler it is to identify objects against a foundation that may not be a lot colder or hotter than the object.

Uncooled Thermal Cameras: These cameras are built on the technology in which the imaging sensor does not require cryogenic cooling. A typical detector model depends on the microbolometer, a modest vanadium oxide resistor with an enormous temperature coefficient on a silicon component with a huge surface region, low heat limit, and great thermal segregation. It works by detecting changes in scene temperature which in turn causes changes in the bolometer temperature. These temperature changes are then converted to electrical signals and finally, they are processed into an image. Uncooled sensors are intended to work in the long wave infrared (LWIR) band, where global temperature targets discharge the greater part of their infrared energy [39]. Comparing the price of uncooled cameras, they are relatively cheaper than cooled thermal camera since it does not require cryocoolers which is expensive. The sensors can be made in fewer strides, with better returns comparative with cooled sensors and more affordable vacuum bundling.

2.3.2 Commercially Available Thermal Cameras

With recent advancements in technology and the evolving market of precision sensors and cameras, many companies are providing commercial systems and solutions which is specifically designed for various thermal imaging applications such as human thermography. The overall system comprises two major parts which include hardware systems and software programs. Currently, third-generation thermal cameras are available in commercial markets which mainly consist of focal plane arrays (FPA) detectors. These detectors use photon detectors also referred to as cooled technology and thermal detectors referred to as uncooled technology. Uncooled detectors are highly preferable for biomedical applications, fire safety, and transportation applications whereas cooled detectors are generally used for high-end military and scientific applications.

A. *FLIR Systems:* It is one of the largest commercial manufacturers specializing in the design and production of thermal imaging cameras and imaging sensors for different commercial and government applications. The company was established in 1978 in the United States of America (USA) and took its name from the acronym forward looking camera (FLIR). FLIR infrared cameras are by and large independent camera models that offer exact, quantitative, and top to bottom estimations. FLIR cameras uncover temperature varieties even under 0.02 °C (20 mK) which means the most profound body or skin recognition is exceptionally conceivable in biomedical applications. The high probability of precision accuracy of FLIR cameras in biomedicine, it provides medical researchers and analysts to extract an increasingly and extensive informative outcome [40-41].

B. *Seek Thermal*: Seek thermal is one of the leading manufacturers of thermal cameras for commercial applications. The company was founded in the year 2012. The company is well known for its low-cost and high-resolution thermal cameras and thermal cores thus advancing the state-of-the-art military and professional-grade thermal technologies. The company is manufacturing its products and solutions for different commercial and government sectors such as biomedical, firefighting, and law enforcement agencies. SEEK thermal cameras mostly rely on thermal detectors that enable them to record infrared-based thermal pictures, especially in the area of biomedicine and other business exchanges. The company is more focused on designing Smartphone-based Seek thermal cameras that can be easily interconnected with android and other smartphones along with their applications thus transforming a cell phone into a professional thermal camera. In addition to that SEEK thermal cameras are more preferred as they are more prudent than other thermal camera models [42].

C. *Infratec*: Infratec stands among one of the largest suppliers of thermal cameras for different applications and specialized hardware equipment for medical applications for approximately 25 years. The company has been providing thermal imaging technology according to customer requirements assigned with safety and security tasks. The performance scope of the hardware equipment is perfected by expert consultancy by providing continuous staff training [43].

Table 2 provides comprehensive information regarding the differences and similarities of infrared thermal cameras manufactured by different companies.

Table 2: Thermal cameras similarities and differences by different manufacturers

Company	FLIR Infrared Camera	SEEK infrared Camera	Infratec Thermal Camera
Area of Application	Biomedicine, building, scientific research areas	Buildings, biomedicine, outdoors	Biomedicine, commercial
Type of Detector	Thermal and Photon	Thermal	Thermal and Photon
Generation	3rd generation	3rd generation	3rd generation
Infrared Spectrum	3-14um	3-14 um	---
Portability/ Mobility	Standalone and smartphone-based	Smartphone-based	Standalone
Type of Measurement	Quantitative and Qualitative	Qualitative	Quantitative and Qualitative
Power Source	Battery-powered	Works on Cellphone battery	Battery-powered

2.3.3 Infrared Spectrums

The infrared region is situated between the visible and microwave regions of the electromagnetic spectrum as shown in Table 3 [44]. Since hot objects/ bodies radiate energy in the infrared band, it is also referred to as the heat region of the spectrum.

Table 3: Infrared Region Ranges

Energy Wave Number ←						
10^8				10^4	10^1	10^-3
Gamma-ray	X-ray	Ultraviolet	Visible	**Infrared**	Microwaves	Radio waves
Nuclear transition	Internal electronic transitions	Electronic valence transition		Molecular Vibration	Molecular Rotation	Spin orientation Magnetic Field
Wavelength →						

The higher the temperature of an object, the higher will be the spectral radiant energy, or emittance, at all wavelengths and the shorter the predominant or peak wavelength of the emissions. Peak emissions from objects at room temperature occur at 10 µm. The infrared region is usually divided into different wavelength regions which include near infrared (NIR), shortwave infrared (SWIR), mid-wave infrared (MWIR), and long-wave infrared spectrum (LWIR).

Near Infrared (NIR): NIR cameras can detect the wavelengths of light directly adjacent to the visible light spectrum. Unlike conventional thermal cameras, NIR cameras still detect photons like conventional RGB cameras in the visible light spectrum, just at a different wavelength. In the NIR spectrum, there are more detectable photons at night, which makes NIR cameras an important source of imaging for night vision and especially military surveillance operations. It works in infrared wavelength spectrum ranging from 0.75um – 1.4um [45].

Shortwave Infrared (SWIR): SWIR also referred to as shortwave infrared detectors offer unique capabilities that are often corresponding to LWIR and MWIR imaging. A SWIR detector is a type of photodetector, like cooled LWIR or MWIR detector. Unlike LWIR or MWIR imaging, SWIR imaging mainly uses reflected light. This is very similar to visible cameras or the human eye. Therefore, SWIR images are comparable to binary visible images in resolution and detail. It generally works in infrared wavelength spectrum ranging from 1.4um – 3µm and frequency band 100–214 terahertz. These cameras can be effectively used for the day as well as night vision imaging.

Midwave Infrared (MWIR): The MWIR thermal cameras operate in the wavelength spectrum ranging from 3–8 um and frequency band of 37–100 THz. These cameras do not require any external light to capture the image as it has the capability to generate an image from the thermal infrared radiation emitted from the body or the object. The brightness of the object acquired from the MWIR thermal imager depends on two major factors which include the object's temperature and its emissivity which can be described as a physical property of materials that explains how efficiently it radiates. The main goal of MWIR cameras is to acquire high-quality images with detailed temperature information and are commonly used for industrial and military applications.

Longwave Infrared (LWIR): Longwave Infrared thermal cameras are also commonly referred to as uncooled thermal cameras. These cameras are designed to operate in the IR band of the electromagnetic spectrum ranging from 8µm to 14µm also known as the LWIR spectrum. Such types of cameras are commonly used for mid-range applications and has the ability to operate in hot as well cold temperatures depending upon the camera hardware designs. As compared to MWIR thermal LWIR thermal camera has better operational capabilities, especially in foggy weather. Moreover, these cameras are inexpensive as compared to cooled thermal cameras.

2.4. LWIR Prototype Thermal Camera

In this project [1] a specialized uncooled LWIR prototype thermal camera has been developed by Lynred France [46] which is further used for all the work package related applications. The prototype camera embeds the LWIR sensor for collecting data in 640x480 VGA resolution for in-cabin as well as outdoor environmental conditions. It has a focal length of 7.5 mm and F-number of 1.2. Figure 6 shows the images of the thermal camera used in this project. The camera is built using uncooled microbolometer-based technology. It is an affordable and lightweight thermal imaging module specially designed to be integrated with AI-based algorithms to extract useful information. The core benefits of this camera include

- Enabling high image quality with low power consumption
- Compact size dimension thus allowing it easily to be fixed anywhere in the car
- Providing agility of configurations and addressing the median volumes application makers.

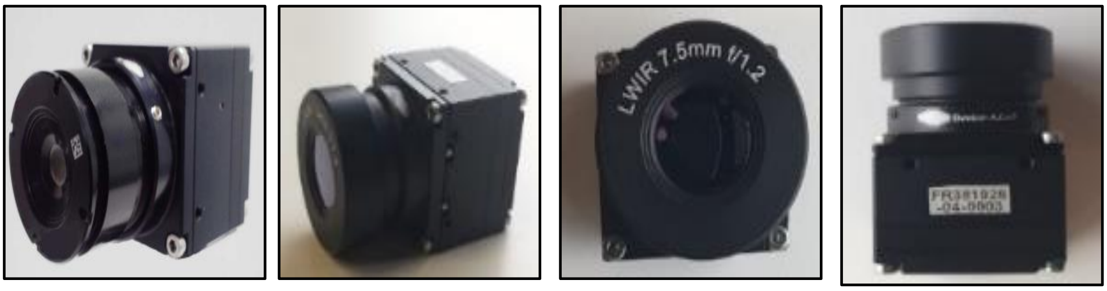


Figure 6: Uncooled LWIR 640X480 prototype thermal camera.

The designed camera module works on plug and plays technology using a USB interface with 16 bits streaming and recording options. Table 4 provides the complete technical details of the thermal imaging camera.

Table 4: Technical details of prototype thermal camera

Specifications		
Uncooled prototype thermal camera	Camera features	Details
Type	Long Wave Infrared	Micro-bolometer technology with 17 µm pixel pitch

Resolution	640 x 480 pixels	
Spectral Response	8 -14 µm	
Focal length	7.5 mm	
Power consumption	< 950 milliwatt	with 30fps recording
Interface	USB	Micro USB type B female, USB 2.0 compliance
Full frame rate	120 hertz	
Standard frame rate	30 hertz	
Export frame rate	9 hertz	
Calibration Methods	Shutterless	With additional image processing algorithms like Bad pixel removal (BPR), non-uniformity correction (NUC), and automatic gain correction (AGC)
Camera Dimensions (Length x Width x Height)	30 x 30 x 24 mm³	Excluding optics and shutter
Camera weight	< 40 grams	

2.4.1 Shutterless Calibration of Uncooled Prototype Thermal Camera

Due to the rapid development of micro and nanotechnology, microbolometers have become significantly inexpensive and more effective. An uncooled IR camera comprises of three main components as depicted in Figure 7.

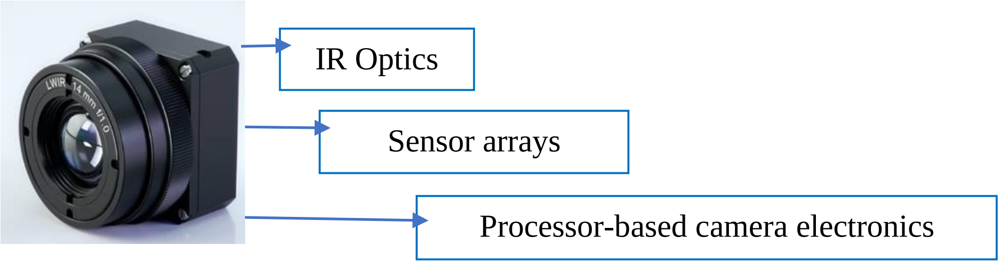
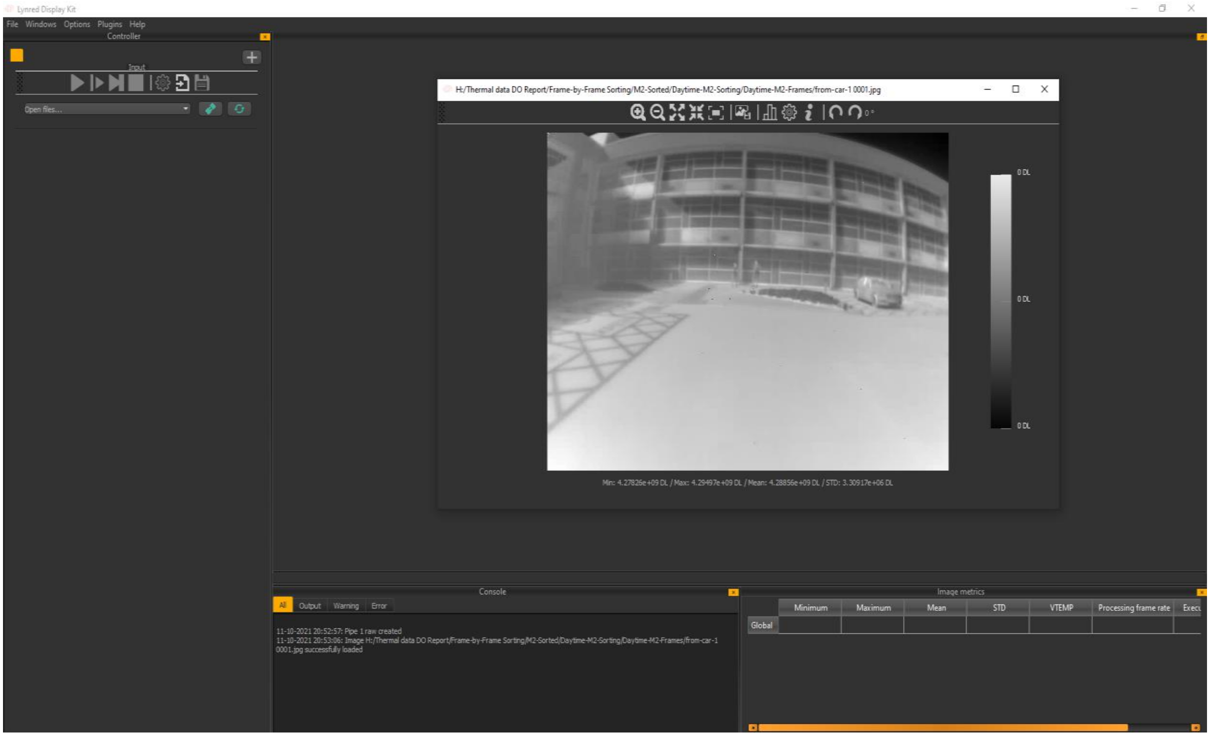


Figure 7: Three core components of an uncooled thermal camera.

These components along with the camera calibration methods play a critical role in ensuring the quality of an IR camera. The calibration is implemented both in the hardware and the software (firmware) [47].

The correction methods for infrared imagers and radiometric cameras vary in the required calibration effort [48]. Radiometrically calibrated thermal cameras typically use optical shutters for runtime re-calibration purposes to regularly correct thermal drift influences on the measurement. The calibration procedure for the shutter-based compensation approach is comprehensively presented in one of the study proposed by Budzier and Gerlach [47]. But the optical shutter is often the size limiting module of an infrared camera since it has to cover the entire aperture. Another drawback of the shutter-based compensation method is the interruption of the measurement during recalibration. Therefore, shutter-less infrared cameras are advantageous, especially for critical real-world applications [49]. Such types of thermal imagers rather than relying on a physical optical shutter to rectify thermal drifts rely on a software approach for image correction. In this method, the shutterless calibration is performed in the camera software/ toolbox.

The prototype thermal camera module is operated through a specialized toolbox built using the camera SDK library. Figure 8 shows the software toolbox GUI for interfacing the camera, loading the plugins, calibrating the camera using shutterless technology, applying various processing operations such as automatic gain correction, bad pixel removal, tone mapping and lastly recording the images or video streams using the Microsoft Windows platform.



H:/Thermal data DO Report/Frame-by-Frame Sorting/M2-Sorted/Daytime-M2-Sorting/Daytime-M2-Frames/from-car-1 0001.jpg

Min: 4.27826e+09 DL / Max: 4.29497e+09 DL / Mean: 4.28856e+09 DL / STD: 3.30917e+06 DL

```
11-10-2021 20:52:57: Pipe 1 raw created
11-10-2021 20:53:06: Image H:/Thermal data DO Report/Frame-by-Frame Sorting/M2-Sorted/Daytime-M2-Sorting/Daytime-M2-Frames/from-car-1
0001.jpg successfully loaded
```

	Minimum	Maximum	Mean	STD	VTEMP	Processing frame rate	Execu
Global							

Figure 8: Uncooled LWIR 640X480 camera display toolbox designed by Lynred France [31] for interfacing the camera, performing camera calibration, and image processing operations.

The initial temperature calibration of the uncooled LWIR prototype thermal camera is done via a specialized black body to provide the hot and cold reference temperature values. A black body is a perfect physical structure that absorbs all incident electromagnetic radiation, regardless of frequency or angle of incidence. The name "black body" is given since it can absorb all colors of light. A black body also emits black-body radiation. Figure 9 shows the black body structure utilized for calibrating the prototype thermal camera.

Shutterless technology allows uncooled IR engines and thermal imaging sensors to continuously operate without the need for a mechanical shutter for Non-Uniformity Correction (NUC) operations. Such type of technology provides proven and effective results in poor visibility conditions ensuring good quality thermal frames in real-time testing situations. For this, we have used a low-cost blackbody source to provide three different constant reference temperatures values referred to as T-ambient1-BB1 (hot uniform scene with a temperature value of 40 degree centigrade), T-ambient1-BB2 (cold uniform scene with the temperature value of 20 degree centigrade), and T-ambient2-BB1 (either hot or cold uniform scene but with different temperature value). The imager can store up to 50 snapshots and select the best uniform temperature scenes for calibration purposes. Once the uniform temperature images are recorded the images are loaded in the camera display toolbox in shutterless confirmation as shown in Figure 10 to finally calibrate the shutterless camera stream

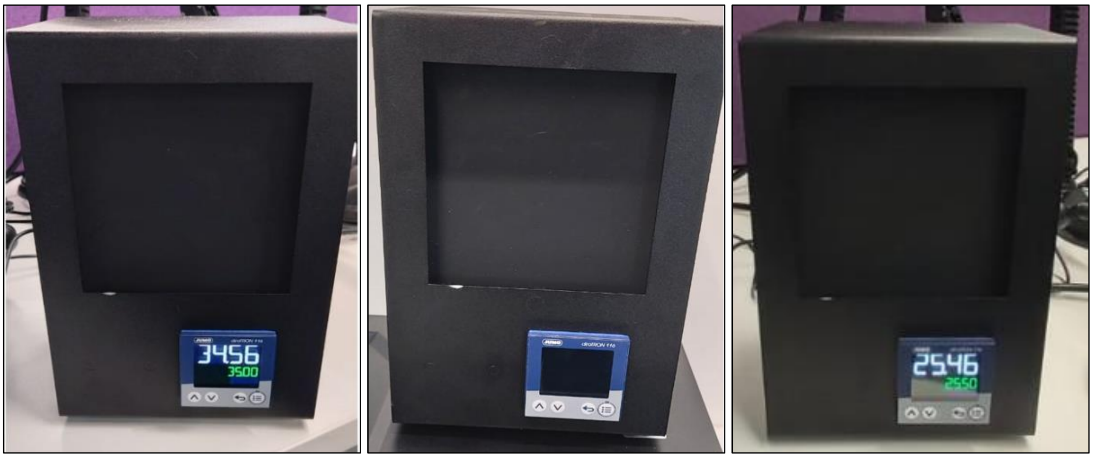
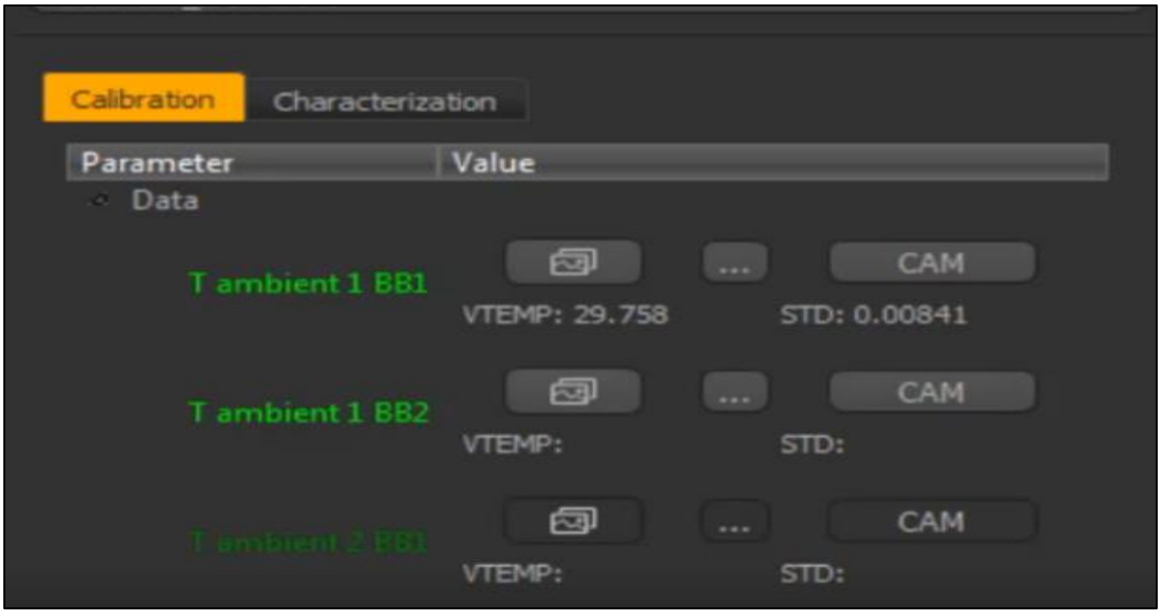


Figure 9: Black body shape for calibrating the thermal camera.



Parameter | Value
Data

T ambient 1 BB1
VTEMP: 29.758 STD: 0.00841

T ambient 1 BB2
VTEMP: STD:

T ambient 2 BB1
VTEMP: STD:

Figure 10: Shuterless camera configuration windows using Lynred display toolbox.

Figure 11 depicts the results of pre and post-processed captured thermal frames using shutterless algorithms on thermal frame capture through the prototype 640x480 thermal IR camera in an outdoor environment. Figure 12 shows the sample thermal frames of two different subjects in the form of six different facial poses acquired through an uncooled prototype camera and using the shutterless algorithm which is enabled through camera GUI as shown in Figure 8 under controlled lighting conditions and in-door environment. It is pertinent to mention that email consent is taken from all the subjects before recording their data. The email prints of these consent forms are attached in the Appendix K of this report.

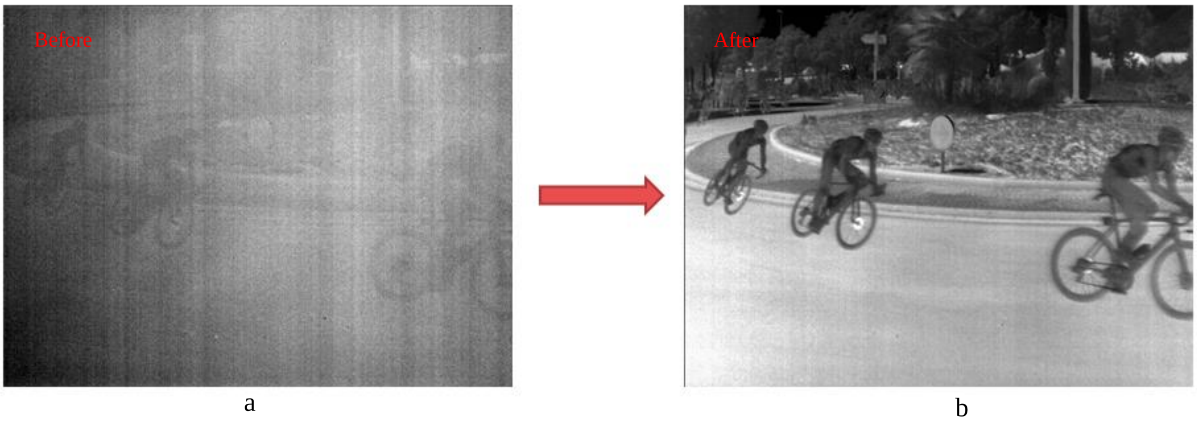


Figure 11: Shutterless calibration algorithm outputs on a sample thermal frame captured from uncooled 640x480 LWIR thermal camera designed by Lynred France [31] a) pre-processed thermal frame, b) processed thermal frame.

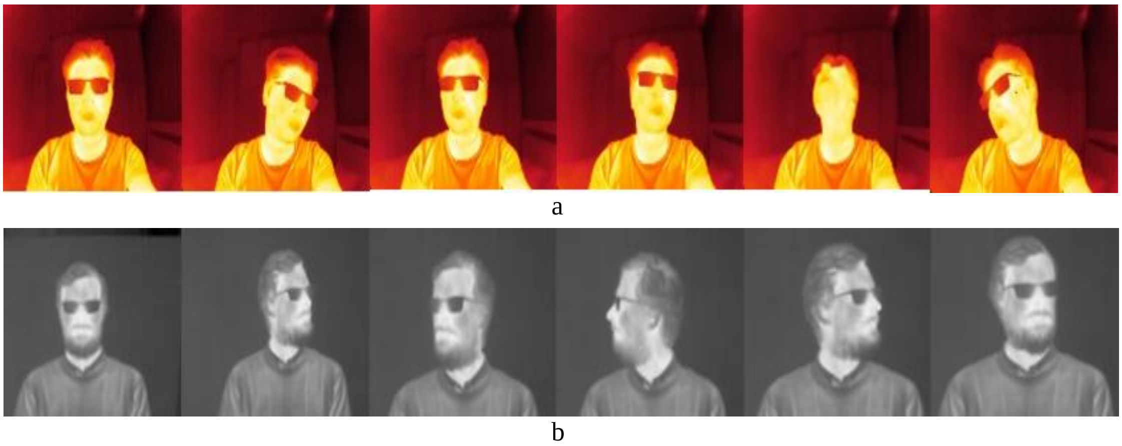


Figure 12: Acquired face data of two different male subjects in 640x480 resolution by applying shutterless technology (a) six different facial poses of a male subject in glow color map, (b) six different facial poses in the grayscale color map.

Chapter 3

Thermal Data Acquisition Methodologies & Regulatory Compliance

This chapter will present complete details about new thermal data acquired locally using the prototype 640x480 uncooled thermal camera. This includes types of thermal data that are being recorded for being utilized in different types of experimental works and data collection methods with complete dataset attributes. The main reason for collecting these datasets beyond publicly available thermal datasets is to complement the requirement of the large-scale thermal dataset and further generalize and validate the CNN networks on locally acquired datasets. The second phase of the chapter will explore the significance of thermal data synthesis that can be used to fulfill the requirement of big data for optimal training of deep learning algorithms. Following this, the chapter will highlight methods and techniques for generating large-scale synthetic thermal datasets and my related publications as contributions towards synthetic thermal data generation.

3.1. Overview of Thermal Data Acquisitions

This section will discuss different types of thermal data that are being recorded/ acquired locally for training and validation purposes while using different types of deep learning models. All these datasets are acquired using the prototype thermal camera developed under the Heliaus project [1] that embeds the Lynred [46] LWIR sensor. The complete camera details and technical specifications of this camera module can be found in chapter 2, section 2.6. During this project, we have collected two different types of datasets which are as follows.

1. Indoor thermal facial data as a part of WP-7 of Heliaus project
2. Out-cabin thermal object detection data as a part of WP-8 of Heliaus project

3.1.1 Indoor Thermal Facial Data

This dataset is comprised of thermal facial images acquired from both male and female subjects who agreed to take part in this research work and provided email consent to record their data. The data is recorded indoor environmental conditions with controlled lighting conditions. This data is useful to carry out empirical work in relation to designing effective in-cabin drive and passenger monitoring systems for enabling enhanced safety features. The subjects were seated at a fixed distance from the camera. The collected dataset is used in various types of experimental work which includes thermal gender classification, and synthetic thermal data generation, using computer vision-based algorithms. It is important to mention that all the data collected during this work is fully in compliance with university data collection rules and regulations. We obtained email consent from all the subjects before collecting their data. All the consent forms are attached in the Appendix K of this thesis report.

3.1.2 Out-cabin Thermal Object Detection Data

This dataset is acquired in outdoor environmental conditions and at different timings such that daytime, evening time, and night-time. The dataset is consisting of several classes of roadside objects such as cars, pedestrians, bikes, bicycles, poles, and buses. This data is useful to carry out pragmatic work in relation to designing effective road monitoring systems for providing comprehensive out-cabin information to the driver. The collected dataset is used in various types of experimental work which includes designing thermal imaging-based object detection and classification system, object detection networks optimization using advanced neural accelerators methods which include TensorRT [51], and deploying the trained networks on resource-constrained edge devices such Nvidia Jetson Nano [52] and Nvidia Jetson Xavier NX embedded boards [53]. It is important to mention that the level of person facial detail, and specific information such as vehicular number plates that can be obtained from the thermal data recorded in the outdoor environment, even in optimal situations, is significantly lower than could be obtained from a typical thermal camera that is optimized for facial detection and recognition which is obligatory under the General Data Protection Regulations (GDPR) and therefore all the data was recorded with GDPR consent obtained from NUIG data protection office. The complete report submitted to the university data protection officer is attached in appendix J of this thesis report. More specifically this report addresses specific concerns regarding the risk to reconstruct facial detail with sufficient resolution to implement a useful facial recognition (FR) and thus to identify individuals and other sensitive information such as vehicular number plates within thermal image data.

3.2. Image Correction Pipeline

This section will explain the adapted image correction pipeline for acquiring high-quality thermal frames for in-cabin and out-cabin vehicular applications. Once the images are refined using shutterless camera calibration as discussed in chapter 2 section 2.4.1 the next stage, employs various real-time image processing-based correction methods to transform the acquired thermal into high-quality thermal frames. This processed data is further used for various computer vision tasks in the context of advanced vehicular applications. Figure 13 shows the adapted three-stage image correction/ processing pipeline.

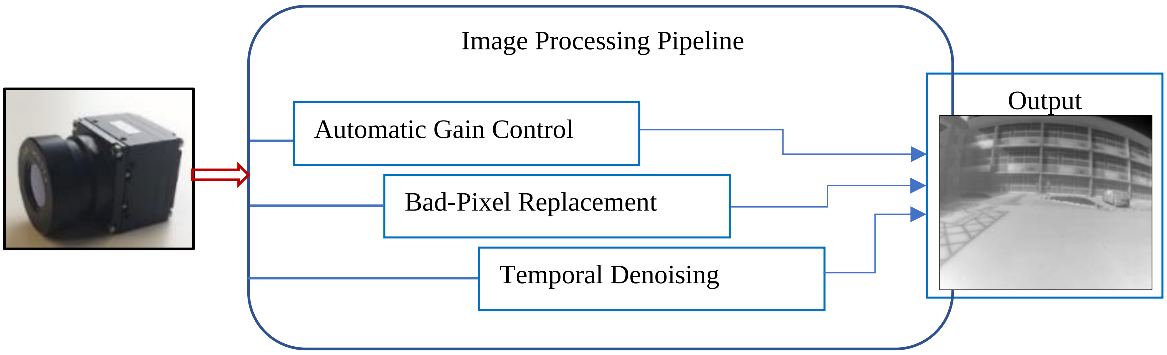


Figure 13: Complete image correction/ processing pipeline to produce high-quality thermal data.

As shown in Figure 13 the whole image processing pipeline consists of three separate image correction techniques which incorporate automatic gain correction (AGC), bad-pixel replacement (BPR), and temporal denoising (TD). The additional details of these methods are provided along with the correction results are provided as follows.

1. Gain Correction/ Automatic Gain Control: The AGC working principle is based on increasing the intensifier gain if the video scene is too dim and decreasing the gain if the video scene is too bright. AGC functions by performing two sets of threshold assessments on the video signal, a level assessment, and a count assessment. First, it level-compares

the magnitude of the signal, pixel-by-pixel, from the camera; a level threshold defines where this comparison is made. Then the AGC performs a count comparison of the accumulated bright pixels (the integrated scene brightness) in each video frame. A count threshold determines when the number of counted pixels defines a bright scene [54].

2. Bad Pixel Replacement: Every pixel in the infrared focal plane arrays (FPA) is distinguished by its offset level, its sensitivity, and its noise level. The pixels that cannot be corrected by the non-uniformity correction (NUC) procedure are usually labeled as bad pixels. Bad pixels rectification is a necessary step as without replacing the bad pixels from the acquired thermal frames can result in poor quality of both the image and the accuracy of the measured data. We can explore several methods which can be used to identify and replace bad pixels, however, the most common and widely used approach is the nearest neighborhood algorithm. This method works by characterizing the signal of the bad pixel which is then replaced by the weighted average of its neighboring pixels. However, the core requirement for this approach to work effectively is that the neighboring pixels should not be the bad pixels. The bad-pixel replacement process is performed by the prototype camera in real-time, after running the shutterless camera calibration [55]. Figure 14 shows the results of BPR algorithms on the sample thermal frame.

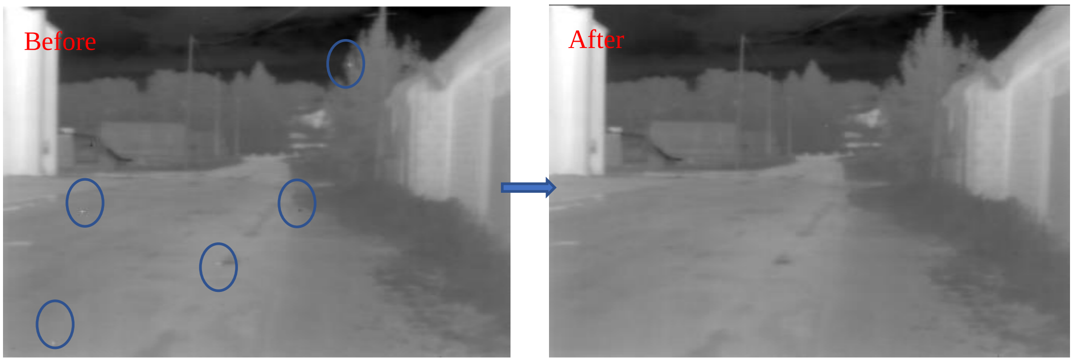


Figure 14: Bad pixel replacement algorithm output on sample thermal frame, left side frame with some bad pixels spotted in blue circles and the right side is processed frame.

3. Temporal Denoising: Temporal noise is a type of random noise that diverges independently from image to image, as compared to fixed-pattern noise, which remains consistent but it is difficult to measure because it is usually much lower than temporal noise. The temporal denoising technique is commonly used to eliminate temporal noise from the thermal frames after the shutterless algorithm is applied.

3.3. Data Collection Methods with Complete Dataset Attributes

This section will describe the complete data acquisition methods along with dataset attributes for both of the locally acquired datasets which include indoor thermal facial data and roadside thermal object detection data.

3.3.1. Indoor Data Collection Methods and Dataset Attributes

This dataset is acquired at two different locations i.e., National University of Ireland Galway (NUIG) and XPERI corporation. The data is acquired in an indoor lab environment using a prototype camera based on an uncooled micro-bolometer thermal array that embeds a Lynred

[46] Long Wave Infrared (LWIR) sensor developed under the Heliaus EU project [1]. The camera is mounted on a tripod stand and the tripod stand is placed at nearly 60 - 65 cm from the subject. The height of the thermal camera on the tripod stand is adjusted physically such that it covers the entire face structure along with the shoulders in the recorded video. Figure 15 shows recorded sample thermal facial sample frames of a male subject during the data recording setup. The complete data recording setup is represented in Figure 16.

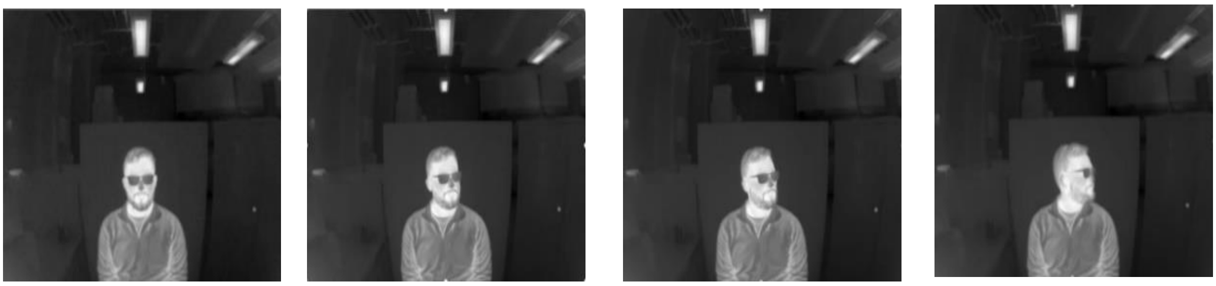


Figure 15: Thermal face data sample frames of a male subject with four different pose variations.

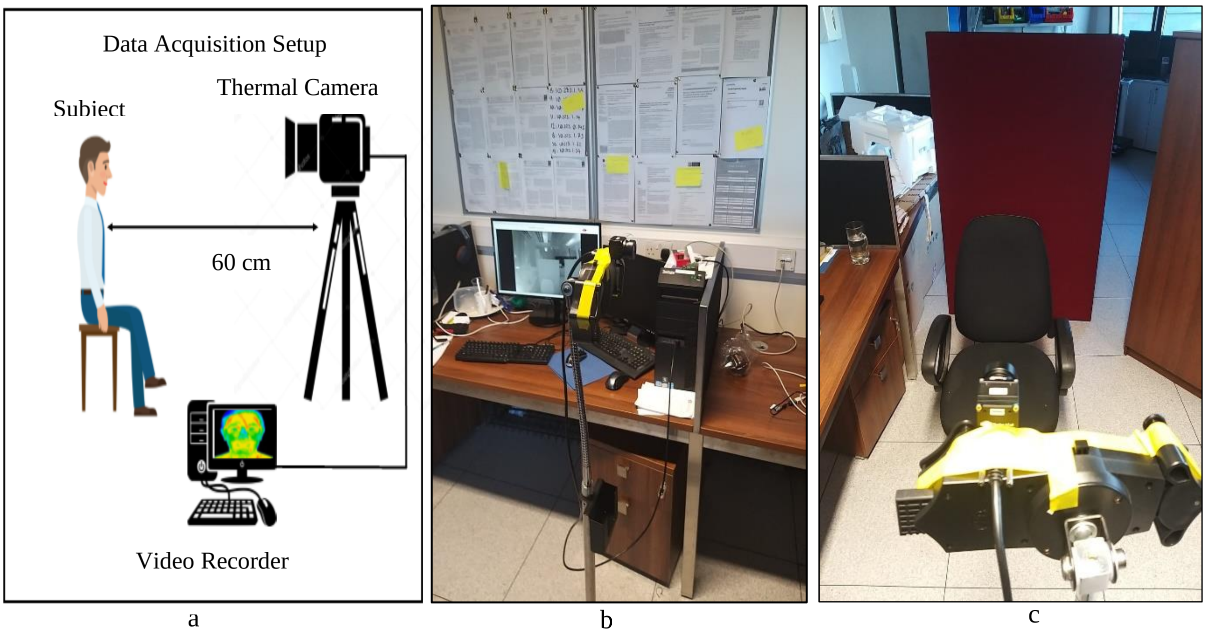


Figure 16: Thermal face data recording setup in an indoor lab environment, a) pictorial representation, b) thermal camera masked with yellow tape on a tripod stand, c) subject seating arrangement with a red background to avoid background clutters.

A total of six subjects took part in this process. The video was recorded at 30 frames per second (FPS) in 640x480 resolution and stored in avi format. The recorded video stream was then converted into facial frames for further experimental work. The video data covers various facial angles to obtain comprehensive facial data. Figure 17 shows the recorded thermal frames extracted from the video files of three different subjects who took part in this study.

In the second phase, thermal face data is collected at the Xperi collaboration lab with slight changes in the overall data recording setup. This data is acquired by integrating a radar sensor in the overall data acquisition setup to record the data at varying distances from the thermal camera with precise distance measurements. For this purpose acconeer radar sensor XM122 [56] is also mounted on the tripod stand along with the thermal camera. The XM122 IoT Module from Acconeer is a low-power connected radar module with an optimized circular form factor with an overall diameter of thirty-three mm. The Acconeer sensor is an mm wavelength pulsed coherent radar, which works by transmitting the radio signals in short pulses which hits the object and rebounces back thus measuring the relative distance of the object from the sensor. It can be used as a separate module where it can be embedded for the customized applications on top of the Acconeer Radar System Software (RSS) using the designed Application Programming Interface (API). It can also be used with an external host controller where different interface protocols can be used such as SPI, I2C, and, UART for communication with the module. The sensor comes with Sofware Development Kit (SDK) for connecting and using the sensor for outside applications. Figure 18 shows the XM122 radar sensor module from the front and back sides.

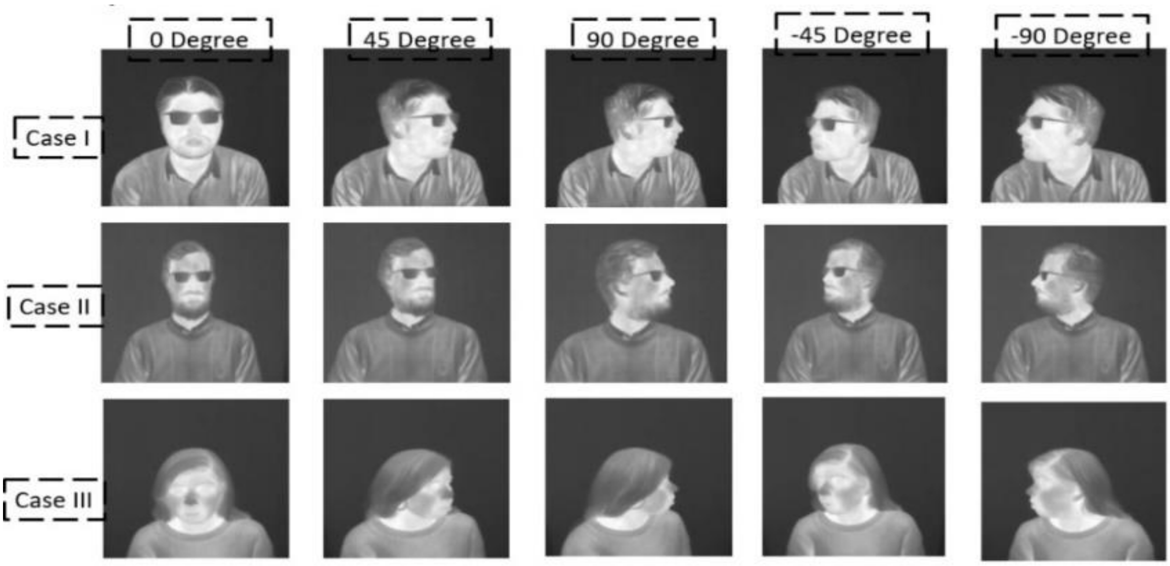


Figure 17: Thermal face data frames of three different subjects acquired in the NUIG lab environment with varying face poses, the first two rows show the five different facial positions of male subjects, and the last row shows the facial angles of a female subject

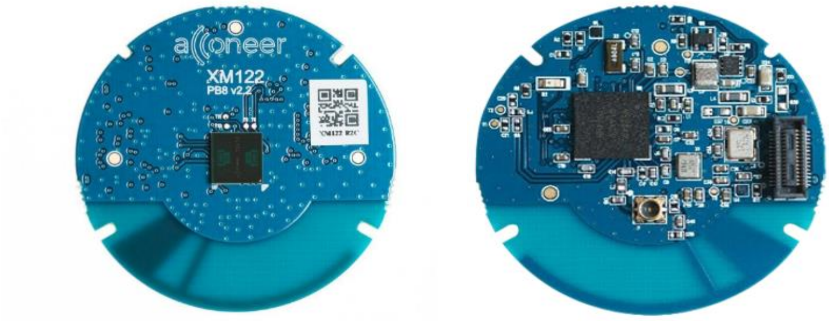


Figure 18: XM122 Aconeer radar sensor module for measuring the object distance.

The key features of this sensor are as follows.

1. The sensor is designed for distance measurement with high precision (millimeter accuracy).
2. It can perform relative measurement with μm accuracy.
3. It can be for various applications such as parking lot sensing, tank level measurement, presence detection, waste bin level measurement, etc.

Figure 19 shows the graphical representation of the data acquisition setup using the aconeer radar sensor.

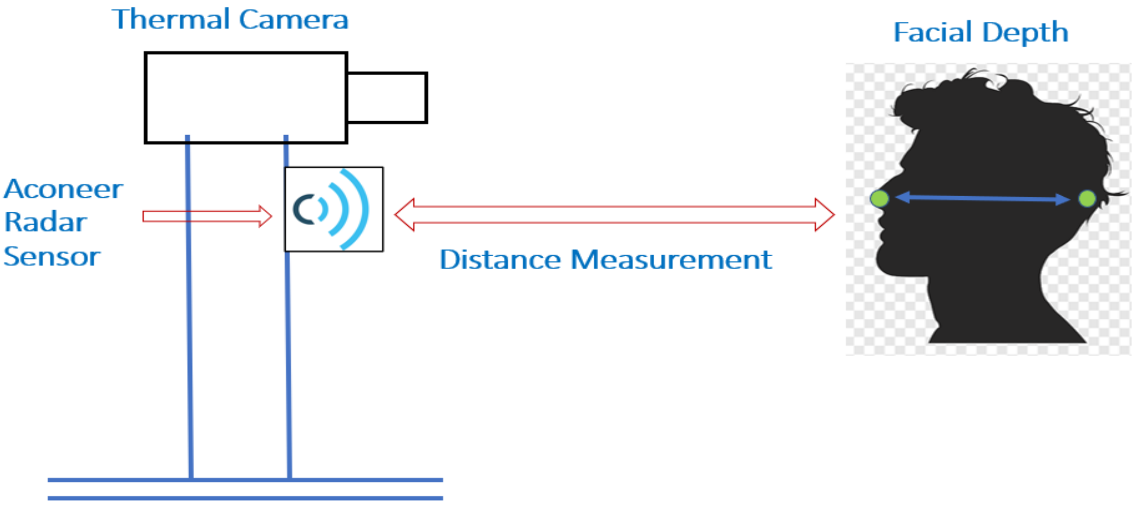


Figure 19: Thermal data acquisition setup using radar sensor and thermal camera.

Before doing the actual data acquisition process some preliminary sensor testings were done by placing the sensor in front of the subject (person) and measuring the real-time distance of the subject from the camera. The measured distance values are cross-validated using the physical measurement scale. The complete setup is depicted in Figure 20.

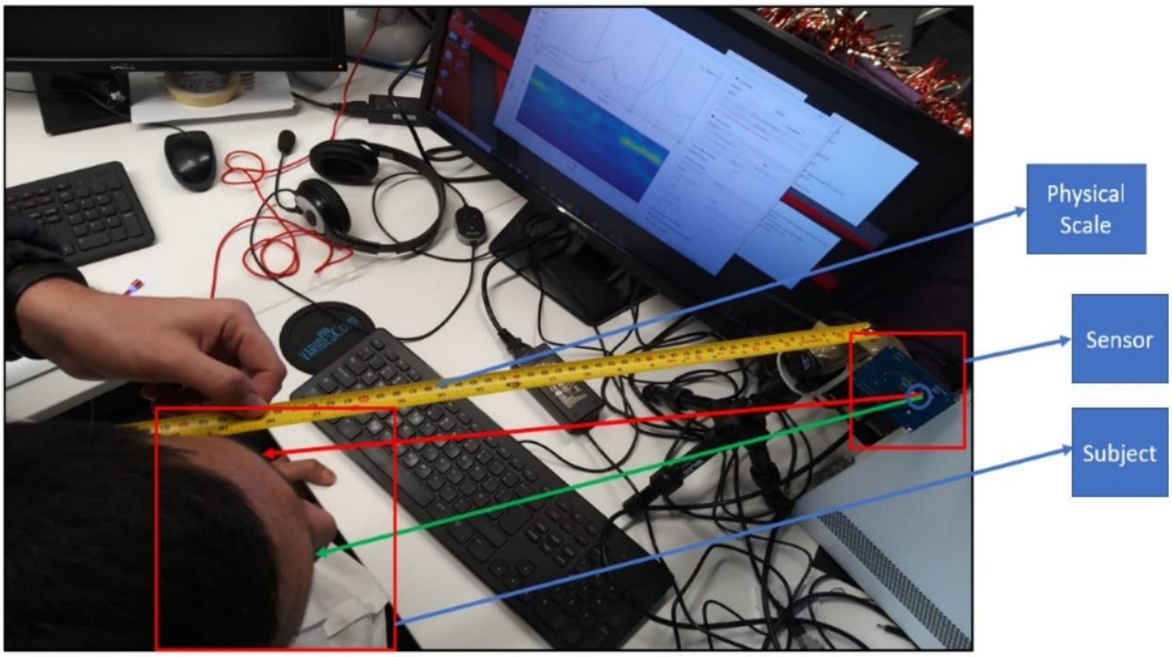
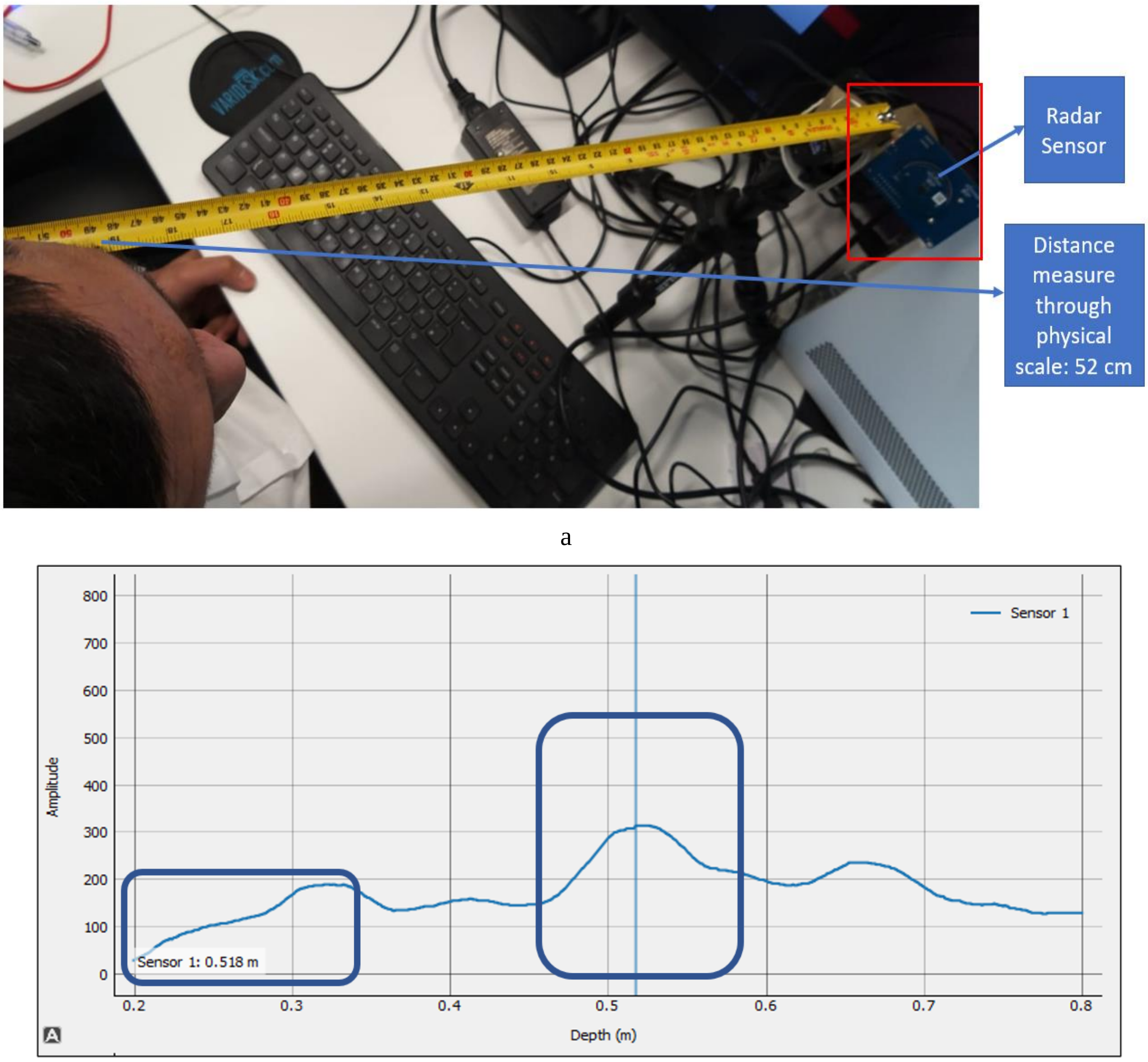


Figure 20: Initial testing of radar module by placing the sensor in from of the subject to compare the distance measured through the sensor and physical scale.

Figure 21 shows the distance measured through the physic scale and results obtained through the sensor module. By comparing the distance values measured through physical scale and sensor module it can be concluded the radar sensor measures distance with good accuracy. After doing the initial sensor testings the next phase is to collect actual thermal data of the subjects at various distances from the camera. The logic behind this type of acquisition is to gather more comprehensive test data keeping in mind the in-cabin driver monitoring and passenger monitoring systems. Driver Monitoring Systems (DMS) make today's car journey safer and more reliable. Because of continued progress in the domain of optical solutions and machine learning algorithms, we can reliably detect some of the major accident-causing factors such as distraction, drowsiness, or even sudden incapacitation. Driver monitoring allows in-time reactions to prevent accidents from happening, therefore it will become a more important requirement for the automotive industry with the passage of time. Also, it plays an important role in the safety of partially automated driving systems. In this data acquisition process, three different distance markings are created from the camera which include 50 cm, 100 cm, and 150 cm. The subjects are asked to stand on those marking points and their actual distance from the camera is determined through the aconeer radar sensor. Figure 22 shows the indoor data collection setup. The overall data is collected using three different marking points from the camera which include 50 meters distance, 100 meter distance, and, 150 meters distance. This is done to complement the requirement for optimal training of DNN networks on locally acquired data for in-cabin vehicular applications.



a

b

Figure 21: Distance measurement comparison using aconeer radar sensor, a) subject distance from the sensor: 52 cm measured through physical scale, b) subject distance from the sensor: 0.518m/ 51.8cm highlighted in the blue box.

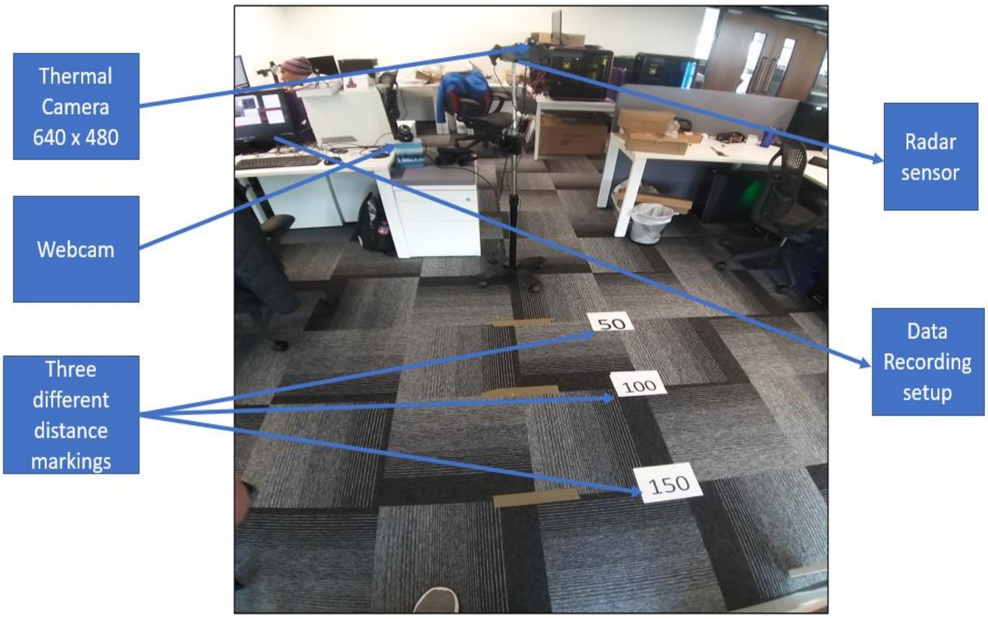
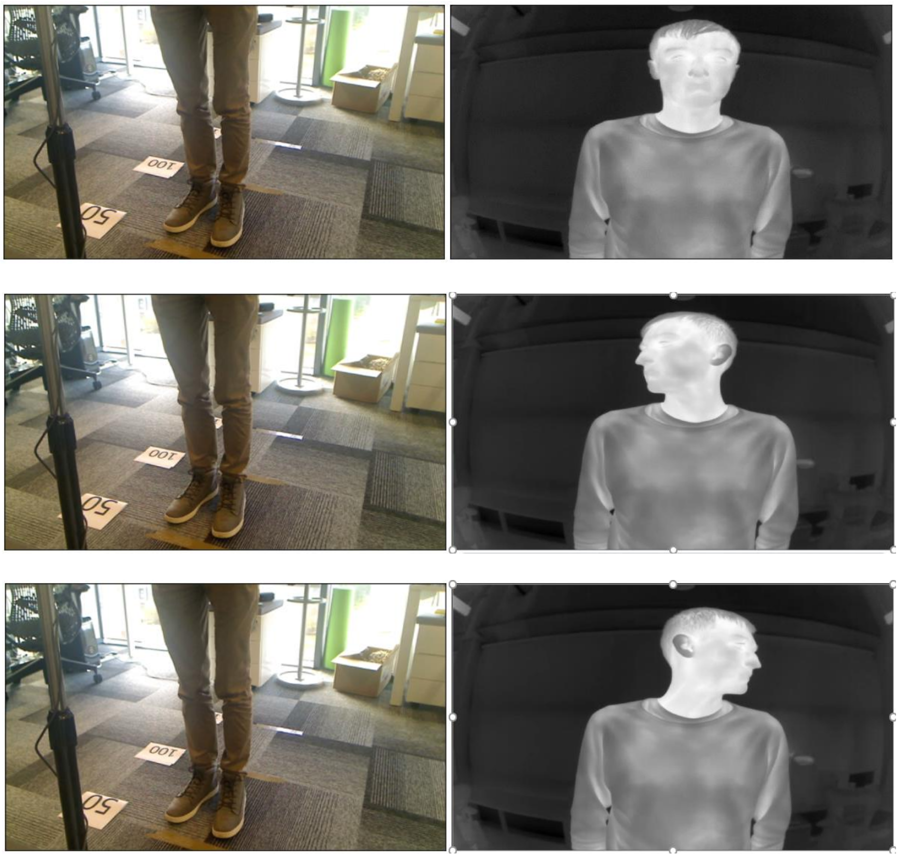


Figure 22: Indoor data acquisition setup in Xperi collaboration lab. The camera and radar sensor are mounted on the tripod stand and an external webcam is used to monitor the subject's movements.

During the data recording process subject were asked to rotate their face from 0 degree to 90 degree right and 90 degree left. Similarly, subjects were asked to move their head in all the direction to capture comprehensive facial poses (yaw, pitch, and roll) for further experimental work. Figure 23 shows the various thermal facial poses of a male subject at a different distances.



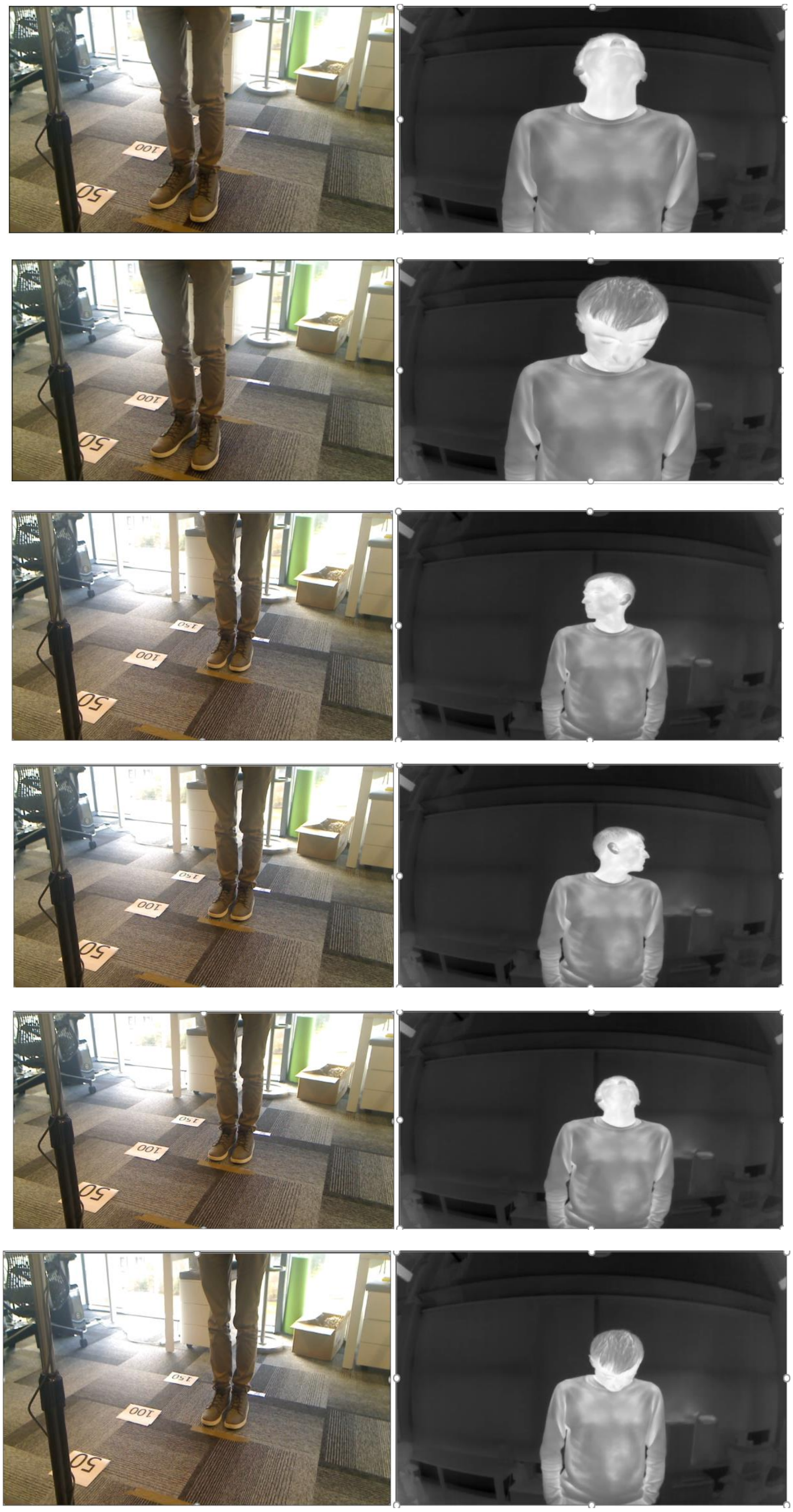


Figure 23: Different thermal facial poses of a male subject acquired from a 640x480 thermal camera. The first five rows show the subject standing at 50cm distance from the camera with five different facial angles and the last four rows show the subject at 100cm distance from the camera with four different facial angles.

Figure 24 shows the aconeer sensor graphical distance reading at one of the data recording instances. It can be observed from Figure 24 that the subject is nearly at a distance of 53.5 cm from the camera using physical markings. For the same point, the sensor reading was 0.538 meters or 53.8 cm which is nearly equal to physical scale readings. Moreover, the subject facial position is 90 degrees right. The same type of data was recorded in the case of all the subjects.

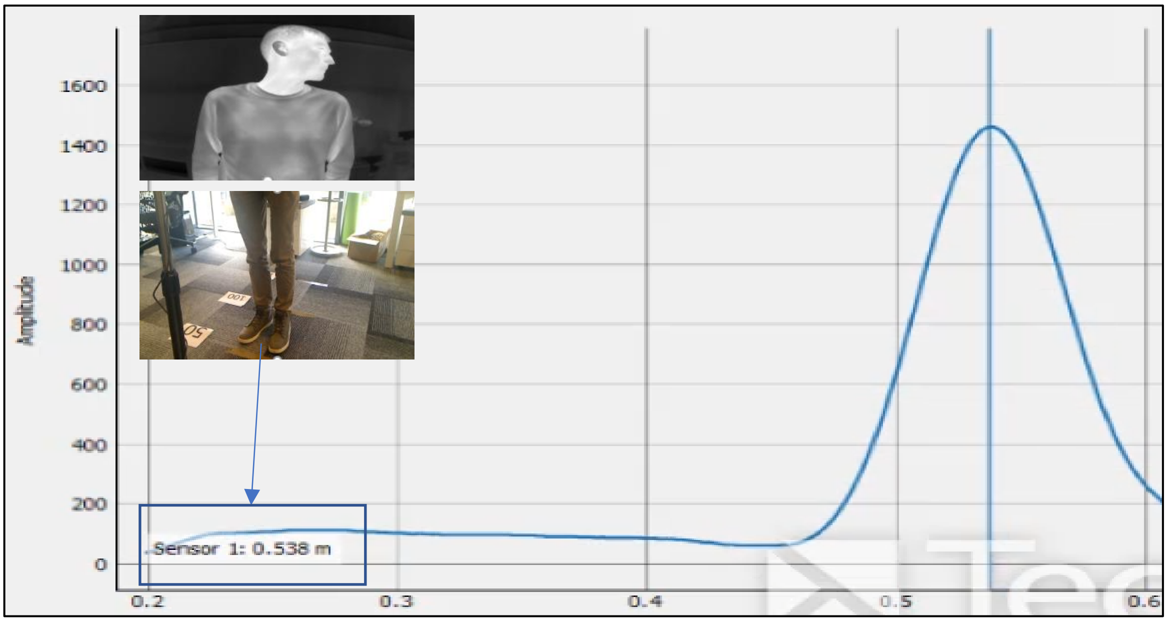


Figure 24: Distance measurement results obtained through the aconeer radar sensor showing the value of 0.538 meters or 53.8 cm of the subject with the maximum amplitude window of 1420.

Table 5 shows the indoor thermal facial data attributes of indoor thermal facial data collected in the NUIG lab environment and Xperi collaboration lab.

Table 5: Indoor Face Dataset Attributes

S. No	Indoor Location	No of Subjects	No of the shortlisted frames	Frames Per Second (FPS)	Key features
1	National University of Ireland Engineering lab	Five	600	30	• includes 5 different face poses
2	Xperi Galway Collaboration lab	Three	400	30	• includes 10 different face poses • Data were collected at three different distance points from the camera

3.3.2. Out-cabin Data Collection Methods and Dataset Attributes

In the next phase, a new thermal dataset is collected in outdoor environmental conditions. This dataset is comprised of six distinct classes which include stationary objects i.e., poles as well as moving class objects which include cars, persons, buses, bikes, and bicycles. The main reason for including all these class objects is that they are most commonly found on the roadside and can be useful for designing effective video analysis-based thermal object detection systems for getting a complete overview of car surroundings. This will eventually help in implementing advanced safety features for intelligent vehicular systems. The dataset is collected in two different methods which are referred to as M-1(By mounting the thermal camera at a fixed place) and M-2 (by mounting the thermal camera on the car) methods. The complete details of these methods along with the comprehensive data acquisition process are provided below. The dataset was collected in the daytime, evening time and, nighttime with challenging weather conditions such as windy weather, fog conditions, and cloudy weather to incorporate sufficient data diversity. The overall dataset was collected in Galway county Ireland.

1. M-1 Method

In, the first approach (M-1) the data is collected in a stand-alone method by placing the camera at a fixed place. The camera is installed on the tripod stand at a fixed height of nearly 30 inches such that the roadsides objects are covered completely in the recorded video. The thermal video stream is recorded at 30 frames per second (FPS). The data is recorded in different weather and environmental conditions. Figure 25 shows the M-1 data acquisition setup for collecting the data from the roadside.

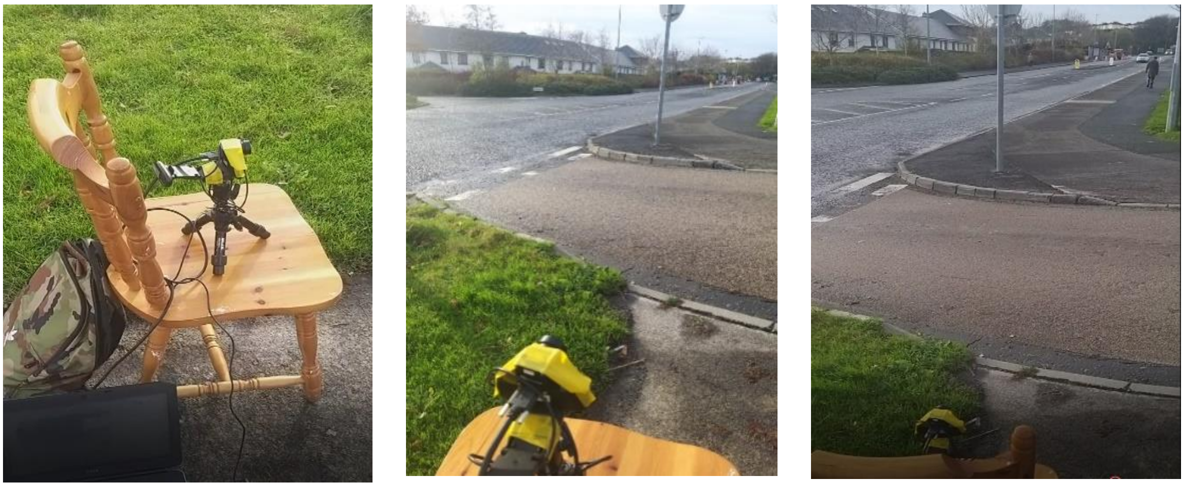


Figure 25: Thermal Data acquisition setup from roadside using M-1 approach such that by mounting the thermal camera at a fixed place.

The overall datasets are collected from roadside and alleyway views in the morning time, evening time, and nighttime conditions.

2. M-2 Method

In the second method (M-2) the data acquisition setup holding the thermal camera and RGB/ visible camera is installed over the electric car and data is recorded in the mobile method. For this, a specialized waterproof camera housing case was printed using a 3D printer to hold the thermal camera in the precise position and angle to cover the entire roadside environment. The housing case also contains a visible camera to get initial visible images as reference data thus allowing us to adjust the thermal camera at an appropriate angle and field of view. Figure 26 shows the camera housing structure and thermal camera video recording setup by placing the camera in the case. After doing the initial video recording testing the housing case is fixed on a suction-based tripod stand thus allowing us to easily fix and remove the complete structure from the car bonnet. Figure 27 shows the camera housing case mounted on the tripod structure and the overall data recording setup fixed on the car bonnet.

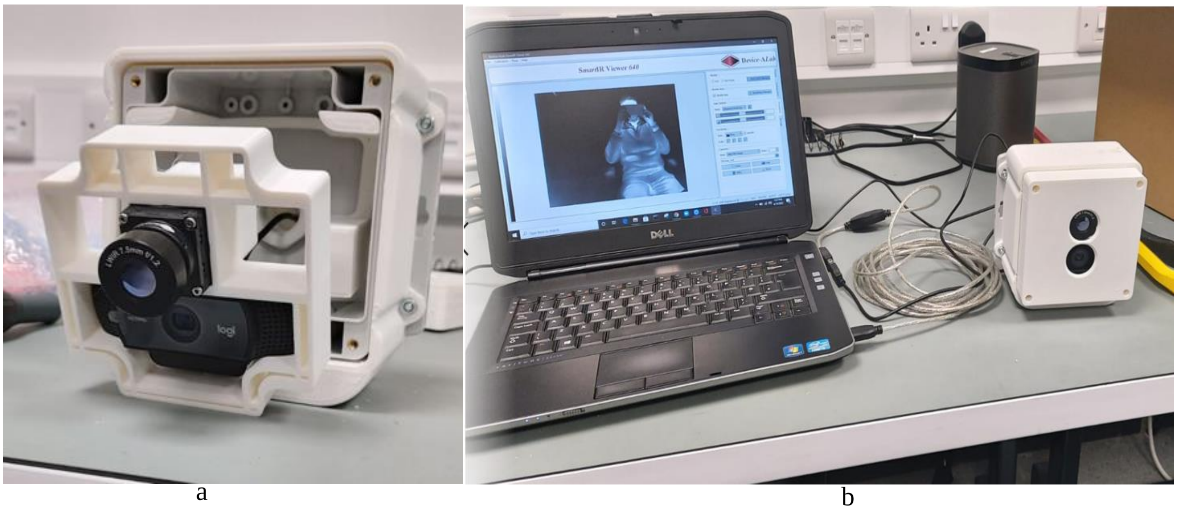


Figure 26: Housing case designed for holding the thermal and visible camera for M-2 method, a) Camera housing structure, b) Initial thermal data recording testings by placing the camera inside the housing structure.

After doing the initial video recording testing the housing case is fixed on a suction-based tripod stand thus allowing us to easily fix and remove the complete structure from the car bonnet. Figure 31 shows the camera housing case mounted on the tripod structure and the overall data recording setup fixed on the car bonnet.

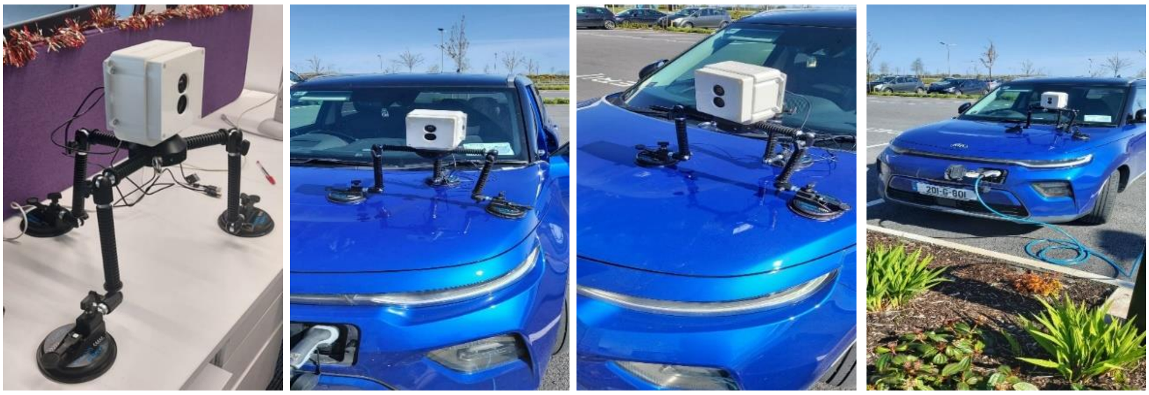


Figure 27: Thermal Data acquisition setup using M-2 method, housing case fixed on a tripod stand and tripod structure is placed on car bonnet with the help of suction cups.

Figure 28 shows the twelve sample thermal frames extracted from the video sets while recording the data using the M-1 and M-2 methods. These thermal frames show the various class objects as discussed in section 3.1.2. The complete dataset is consisting of a total of 6 video sets, 39,770 distinct thermal frames and 2200 ground-truth annotated frames. The complete dataset attributes are summarized in Table 6. The main reason for collecting the data in two different methods is to bring variations and collect distinctive local data in different environmental and weather conditions. This data is further used for robust training of YOLO-v5 end-to-end deep learning architectures to deploy Thermal-YOLO frameworks on embedded GPU hardware for enabling enhanced video analysis-based safety features for driver assistance. The complete dataset is published and open-sourced through IEEE Dataport (Link: https://ieee-dataport.org/documents/c3i-thermal-automotive-dataset) [17].

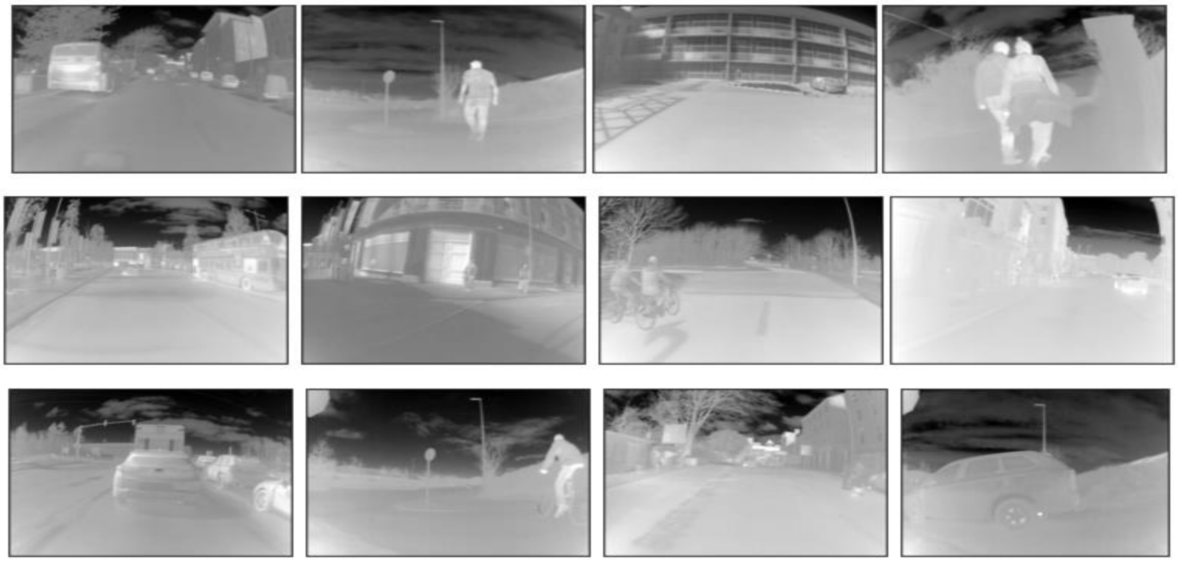


Figure 28: Twelve distinct thermal frames were captured using LWIR 640X480 prototype thermal camera using M-1 and M-2 methods.

Table 6: Outdoor thermal dataset attributes

Locally acquired dataset attributes				
Data collection method with frame properties	Total number of extracted frames	Processing Method	Environment	Time and weather conditions
M-1 Camera mounted at a fixed place 96 dpi (horizontal and vertical resolution) with 640x480 image dimension	8,140	Shutterless, AGC, BPR, TD	Roadside	Daytime with cloudy weather
	680		Alleyway	Evening time cloudy weather
	4,790		Roadside	Night-time with light cloudy and windy weather
M-2 Camera mounted on the car (Driving condition) 96 dpi (horizontal and vertical resolution) with	9,600	Shutterless, AGC, BPR, TD	Industrial Park	Daytime with clear weather and light foggy weather
	11,960		Downtown	Evening time with partially cloudy and windy weather

640x480 image dimension	4,600	Shutterless, AGC, BPR, & TD	Downtown	Night-time with clear weather conditions
frames	*Daytime: 17,740 (44.61%)*	*Evening time: 12,640 (31.78%)*	*Night-time: 9,390 (23.61%)*	*Total: 39,770*

3.4. Contribution to In-Cabin Data Acquisition By Xperi

In addition to collecting our thermal datasets as discussed in section 3.3.1 and section 3.3.2, I also contributed by helping the team of engineers in Xperi for developing data tool that synchronizes and reads multimodal data in a single GUI to further used for in-cabin applications. Moreover, I also participated as a volunteer subject in Xperi in-cabin data collection. Xperi was undertaking very substantial data and the goal of this data was to observe natural drowsiness behavior and high cognitive load in a simulated driving situation with several optical and electrical sensing modalities. The main purpose of this extensive data is to further use it to help in improving driver safety by enhancing the driver and occupant monitoring systems by including robust drowsiness and cognitive load detection and prediction capabilities. As mentioned the acquired dataset was to be used for in-cabin applications, therefore rather than collecting such type of extensive data in the university environment and due to certain covid restrictions at that time, I supported the data acquisition team in Xperi by providing technical assistance and also participating as volunteer subject. Figure 29 shows the overall data acquisition setup at the Xperi Galway office.

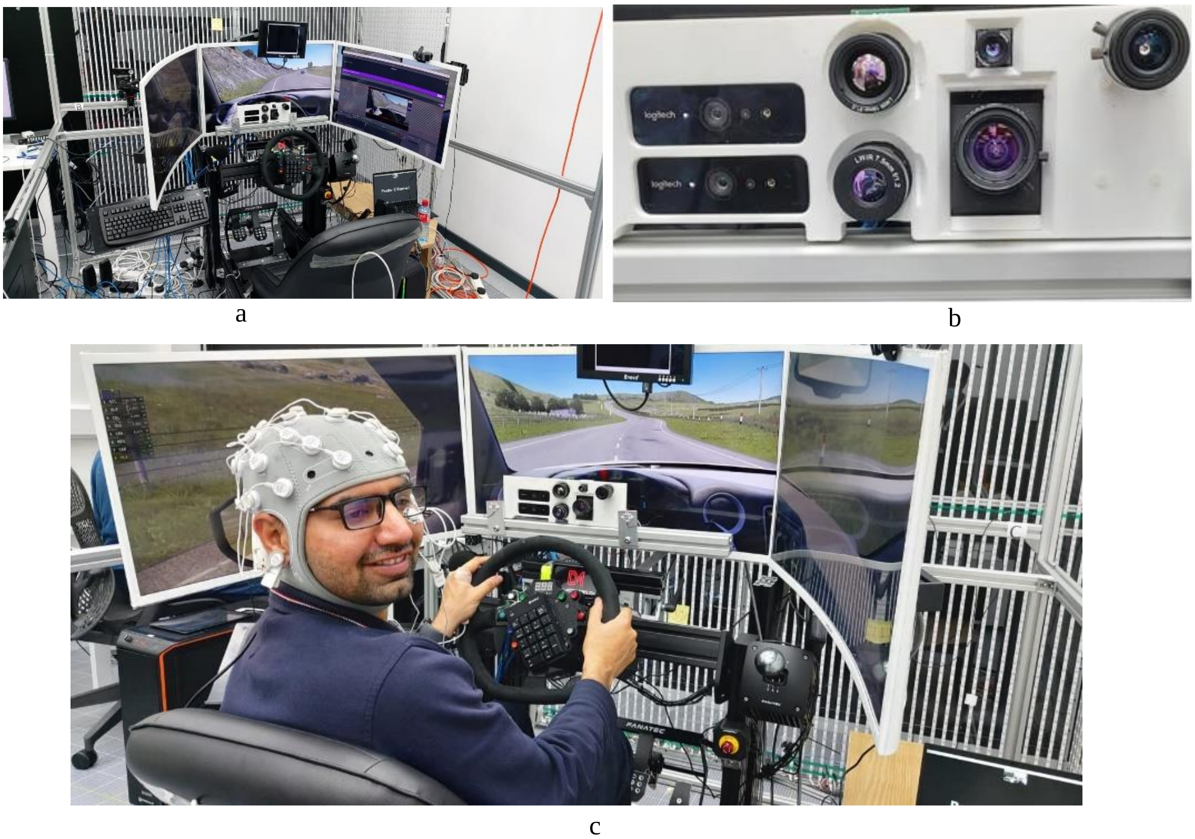


Figure 29: Data acqsation setup for drowsiness and cognitive load monitoring at XPERI Galway, a) Driving simulator setup with three wide angle monitor mounted for having real world driving experience, b) Different types of image sensors mounted at the center of middle display to record driver facial data, c) My self seated in the driver simulator wearing EEG, EOG, EDA, ECG, and SpO2 sensor kits.

The proposed research involves approximately 5 hours in a driving simulator (with breaks) while the volunteer sits in a driving simulator and is asked to perform various tasks while the data is recorded with several sensing modalities. These sensing modalities include EEG, EOG, EDA, ECG, SpO2, NIR, RGB, Audio, and Thermal IR. The foremost purpose of including the joint fusion of various electronic and optical sensors is to collect and further analyze the electrical activity data from the brain, heart rate data, data about eye muscle movements, data on the perspiration around the skin of the wrist, and temperature-related information.

3.5. Investigations in Data Synthesis, Data Augmentation & Generative methods

Deep Neural Networks (DNN), such as Convolutional Neural Networks (CNN), are bridging the gap of automation and have made an incredible improvement in discriminative tasks, but it needs lots of training data for achieving optimal training results and robust validation results in many computer-vision applications. The state-of-the-art pretrained architectures are requiring substantial volumes of training data such as annotated data for the training of object detection models. However, such an approach is costly, prone to errors, and labor as well as time-intensive, especially in highly complex, dynamic production and real-time environments. To overcome this barrier, synthetic data can be created by using suitable training datasets as seed data and utilized for accelerating the training phase of DL. Synthetic data is a type of data that is generated artificially instead of being generated by actual events. The generation of synthetic data is far cheaper compared to data acquisition and in many cases, these data samples come with annotated labels. It is often generated with the help of computer vision algorithms and is used for a wide range of activities for instance generating new test data for validating the performance of trained architectures, and in deep learning models tuning for specific applications. It is difficult to find large-scale datasets in thermal imaging modality therefore synthetic data plays a pivotal role at this point for optimal generalization of deep learning architectures. Finally, but equally importantly, since this data is generated using artificial methods, there are no underlying GDPR/privacy concerns and it can be used freely. After an extensive study of various methods for generating synthetic data, we have concentrated on three different methods for generating synthetic thermal data using the existing thermal datasets. These methods are as follows.

1. Data augmentation or data transformation

2. Generating fake thermal data using StyleGAN (Generative Adversarial Networks)

3. 2D to 3D face reconstruction using end-to-end PRNet (deep learning networks)

The generated synthetic data using these methods were used for the training purposes of pretrained CNN architectures for thermal gender classification. The performance of CNN models was presented, compared, and discussed in one of our published study [9]. The results indicate that CNNs trained on synthetically generated datasets have acceptable performance as compared to the model trained on raw thermal data.

3.5.1. Data Augmentation

To develop an effective deep learning model, the validation error must continue to reduce with the training error as well. Data Augmentation is a very powerful method for achieving this goal. The augmented data will correspond to a more comprehensive set of possible data points, thus minimizing the distance between the training and validation set, as well as any future testing sets for further validation of trained DNN. Data augmentation can be defined as the method to increase the existing data by adding new copies of data samples with slight modifications thus producing the data diversity [57]. This approach helps in preventing overfitting in the DNN [57]. In this work [9], we have mainly focused and implemented various geometric data transformation methods (Appendix A) on exiting thermal datasets as the seed data to generate modified data samples.

3.5.2. Generating fake thermal data using StyleGAN (Generative Adversarial Networks)

Generative Adversarial Networks (GANs) is a method of generative modeling that is based on deep learning methods such as CNN. It works in an unsupervised learning fashion by learning the patterns in the input data such that the trained network can be used to generate new data samples. The main concept of GAN architecture was first presented in 2014 by Ian Goodfellow, et al. titled as “Generative Adversarial Networks” [58]. In this work [9], StyleGAN [59] has been employed for generating synthesized thermal facial samples (Appendix A) by using the existing thermal facial datasets as seed data. StyleGAN is state of art GAN network introduced by NVIDIA researchers having the capability to generate seemingly vast numbers of high-resolution data samples. In our experimental work, the styleGAN network was trained on a variety of thermal datasets which include the tufts dataset, carl dataset, and Laval face motion thermal datasets. The complete experimental details regarding the training and testing approaches of styleGAN are published in our conference papers titled as “Proof-of-Concept Techniques for Generating Synthetic Thermal Facial Data for Training of Deep Learning Models” at the 39th International Conference of Consumer Electronics (ICCE 2021) [9].

3.5.3. 2D to 3D face Reconstruction using End-to-End PRNet

This method is used to generate multiple 3D facial geometric structures using the single 2D thermal facial frame. This is achieved by using a pretrained end-to-end convolution net referred to as Position Map Regression Network (PRNet) [60]. The overall network structure functions by transferring the input image into a position map. In the second step, the encoder-decoder method is employed for learning the transfer structure. The encoder block comprises of single convolution layer which is followed by a series of ten residual blocks for performing downsampling operations on the data. The decoder block comprises seventeen transposed convolutions blocks to generate the predicted output position map. PRNet CNN uses Rectified linear Unit (ReLU) activation functions and a kernel size of four is applied for each of the convolution layers and transposed convolution layers. Figure 30 shows the complete workflow diagram from our work [10] for generating synthesized 3D facial geometric structures from single 2D thermal frames.

The complete experimental details for generating the synthesized 3D facial structures using PRNet is published in our conference papers titled as “Generating Thermal Image Data Samples using 3D Facial Modelling Techniques and Deep Learning Methodologies” in the 12th International Conference on Quality of Multimedia Experience (QoMEX 2020) [10]. The copy of the paper published based on this section is presented in Appendix B.

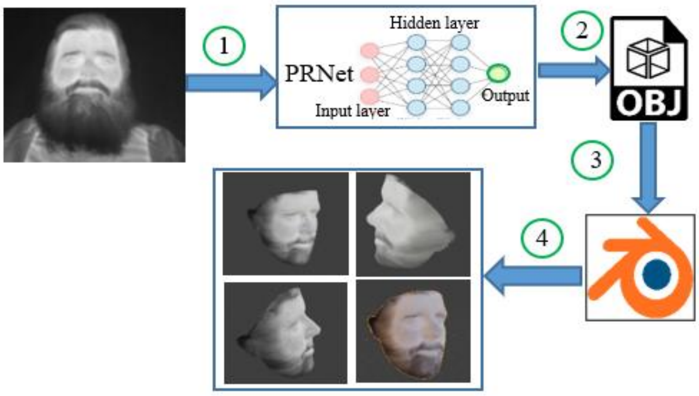


Figure 30: Comprehensive workflow diagram for generating the synthetic 3D facial structure from 2D thermal image 1: input images fed to PRNet for generating 3D facial geometry data, 2: PRNet outputs an obj file, 3: obj file is imported to blender software, 4: final outputs extracted in the form of 3D thermal facial images covering different facial angles and poses.

3.6. Conclusion on Data Acquisition & Synthesis in Thermal Imaging

This section will highlight the focal applications of acquired thermal and publicly available thermal datasets, along with the artificially generated synthetic thermal data in various experimental work carried out under this project. For outdoor thermal data, we have obtained approval from the university data protection officer. The locally acquired data along with the public datasets is used in the diversified application for in-cabin driver monitoring as well as out-cabin monitoring systems. These applications are as follows.

1. Thermal gender classification system

The further details of our contributions related to the thermal gender classification system are presented in Chapter 4 and the complete published study is attached in Appendix C of this thesis report.

2. Object detection in thermal spectrum for ADAS

The further details of our contributions related to object detection in thermal spectrum are presented in Chapter 5 and the complete published study is attached in Appendix E of this thesis report.

3. Object detection algorithm optimization for deployment on Embedded GPU devices.

The further details of our contributions related to object detection algorithm optimization for deployment on embedded GPU devices are presented in Chapter 6 and the complete published study is attached in Appendix F of this thesis report.

4. Face localization and facial landmarks detection in thermal images

The further details of face localization and facial landmarks detection in thermal images are presented in additional experimental work which is attached in Appendix L of this thesis report.

Chapter 4

Contribution To Development of In-Cabin Thermal Gender Classification System using SoA CNNs

Gender classification found many useful applications in the broader domain of computer vision systems such as in-cabin driver and occupant monitoring for autonomous vehicles, smart surveillance systems, people counting especially in crowded areas, etc. This chapter will focus on core contributions towards the development of an efficient thermal gender classification system using end-to-end convolution neural networks for in-cabin driver monitoring applications. This was our first extensive experimental work carried out under WP-7 of the Heliaus project using neural classification algorithms. The goal of this work is to develop a thermal soft biometrics classification framework, compliment the behavioral work of NEXT2U, and the facial analytics and super image resolution work of Xperi for WP-7.

4.1. Research Objectives

The Heliaus project [1] focuses on the sense and think part of the perception process for both in and out-of-cabin applications. Therefore acquiring human body details is very necessary nowadays, especially for human-computer interaction systems, where the machine needs to classify the person's characteristics using various types of data. It includes face recognition, gender classification, facial expressions recognition, and drowsiness detection. The human gender classifications system is a crucial requirement for many critical systems such as autonomous vehicles, and smart surveillance systems deployed at public places such as airports and railways stations, shopping centers, government buildings, etc. The most important applications include smart driver monitoring systems (DMS) for autonomous vehicles where the system acquires useful information about the driver to configure vehicle responses and configuration to ensure maximum comfort and vehicle safety. For instance, it can be used to better predict driver cognitive response, driver behavior, and intent, and finally knowledge of gender can be useful for safety systems such as airbag deployment that may adapt to driver physiology.

We can find various algorithms in the fields of computer vision and machine learning for efficient gender classification. Previous studies have proposed many different methods for the gender classifications system which can be divided into two main categories voice data [61] and video/ image data [62]. The joint approach of both methods can lead towards better accuracy levels, however, the noise generated through analog sensing devices like microphones greatly affects the results. Conventional machine learning algorithms like Support Vector Machines (SVM) are used for different types of biometric classification systems such as fingerprint identification, face recognition, gender classification but the main drawback of these classifiers is that it depends on manual feature engineering process (hand crafted features) using various algorithms which mainly includes PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis). PCA performs linear operations to create new features. PCA fails when the data is non-linear and is not able to create the hyperplane or decision boundary between different classes [63] thus the effectiveness/ accuracy of traditional machine learning algorithms is significantly affected if the features are not extracted precisely. Moreover, the performance of SVM will underperform and drops sharply in cases where the total number of feature values for each data point exceeds the number of training data samples and also when the provided data is noisy such that target classes are overlapping with each other. Thus, to overcome such types of challenges we have typically focused on Deep Neural Networks (DNN) as it plays an imperative role in achieving more accurate and robust results. In recent years, Convolutional Neural Networks (CNNs) have become the state-of-the-art for object classification/ recognition and detection tasks in the domain of computer vision for diversified real-world applications. Typically, a CNN structure consists of several convolutional layers, max-pooling layers followed by fully-connected (FC) layers. As compared to traditional machine learning algorithms such as support vector machines (SVM) which mainly relies on manual feature engineering process the CNN can self-extract the feature values also referred to as automated feature extraction using raw pixel values from the provided image/ video data. Indeed, the real quality of deep learning models comes from an extensive feature engineering process than from the modelling technique itself [64]. While specific machine learning techniques may work best for tasks (problem/dataset), features are the universal drivers/critical components for any modelling application. Extracting as much information as possible from the available datasets is crucial to creating an effective solution. The second most important factor is the training data and fine-tuning process of traditional and deep learning models. Support vector machines effectively use only a subset of a dataset as training data. This is because they reliably identify the decision boundary on the basis of the sole support vectors. Therefore, for well-separated classes, the number of observations required to train an SVM isn't high. With regards to convolution neural networks, instead, the training takes place based on the batches of data that feed into it. This means that the specific decision boundary that the neural network learns is highly dependent on the order in which the batches of data are presented to it. This, in turn, requires processing the whole training dataset without the need to break the datasets into smaller subsets. Moreover, the rapid developments in the world of deep learning, image classification, detection, and segmentation has been further accelerated by the advent of the transfer learning technique. Transfer learning allows us to use pretrained models such as Inception-v3 [65], and EfficientNet [66] which are already trained on a big dataset, we can further use these models for custom tasks. Consequently, reducing the cost of training new deep learning models and since the datasets have been vetted, we can achieve precision accuracy and optimal generalization of the model with reduced computation cost and lesser training time. Once the networks are trained by the appropriate splitting of datasets and selecting proper loss function, generalizations, and optimization techniques, the trained DNN models can be deployed and used rapidly to predict the results on unseen test data.

Further, we have highlighted some of the published studies to compare the performance of CNN over SVM for different applications and their respected outcomes. In one of the recent study titled "Critical Comparison of the Classification Ability of Deep Convolutional Neural Network Frameworks with Support Vector Machine Techniques in the Image Classification Process" by Robert Kelly [67], the performance of CNN and SVM classifiers is evaluated on a dataset of approx. 55,000 images. This dataset was used to assess the classification potential of each methodology, in terms of training, implementation, and the ability to engineer parameters and features for successful classifications on a very large dataset. The individual performance of each of these methods was compared using different parameters which include, training time, confusion matrix, and ease of use to assess which has the higher classification potential. The overall outcomes in the form of various experimental hypotheses indicate that the application of deep learning techniques had an adequate edge over SVM approaches in both accuracy and data handling, that to not natively avail of the computational power of deep learning models. In another study [68] authors have compared the performance of three different machine learning methods which incorporates CNN, ANN, and SVM for medical image analysis using CT, MR, and X-ray imaging datasets. The experimental results concluded that CNN exhibited higher accuracy as compared to SVM and ANN methods. Moreover, the study further summarizes that CNN has a significant advantage as it exhibits a key attribute to accurately identifying clinical images with less amount of data and in a shorter period; thus, incorporating CNN into computer-aided medical image processing and inspection systems is beneficial.

In this work, we have developed an AI-based autonomous thermal gender classification system using a set of pretrained CNN architectures and proposing a novel CNN architecture referred to as 'GENNET'. This work belongs to work package 7 (WP-7) of the Heliaus project [1] which is focused on the development and validation methodologies of the thermal-IR system for in-cabin vehicular applications.

4.2. Summary and Discussions of Contributions

In this work, diverse thermal and RGB public datasets had being utilized which include Tufts thermal face database, Casia Dataset, and Carl thermal dataset for training and validation purposes of deep learning architectures. The main contributions of this work are presented in bullet form.

- In the first phase, we have trained nine state-of-the-art pre-trained networks from scratch (by unfreezing all the network layers) on a large-scale casia facial dataset [69]. These models includes AlexNet, VGG-19, MobileNet-v2, Inception-v3, ResNet-52, ResNet-50, ResNet-101, DenseNet-121, Dense-201 and EfficientNet-B4. The trained architectures are further fine-tuned using Tufts public thermal dataset [12-14].

- In addition to employing the pretrained architectures [11], the main contribution of this work is designing a novel CNN architecture 'GENNet' for the thermal gender classification task, and further its performance is compared against all the pre-trained state-of-the-art architectures. The structural block diagram representation of the GENNET architecture is shown in Figure 31.

- In the second phase, the overall efficacy of a wide range of pretrained CNN along with newly proposed GENNet architectures is validated on the combination of two different test datasets which includes Carl thermal public dataset [15-16] and newly acquired thermal data in the NUIG indoor lab environment. The complete details of the test set can be found in (Appendix C).

- For rigorous validation tests of all the trained architectures on thermal data we have used nine different quantitative metrics. These include accuracy, sensitivity, specificity, precision, negative predictive value, False Positive Rate (FPR), False Negative Rate (FNR), Matthews Correlation Coefficient (MCC), and F1-score.

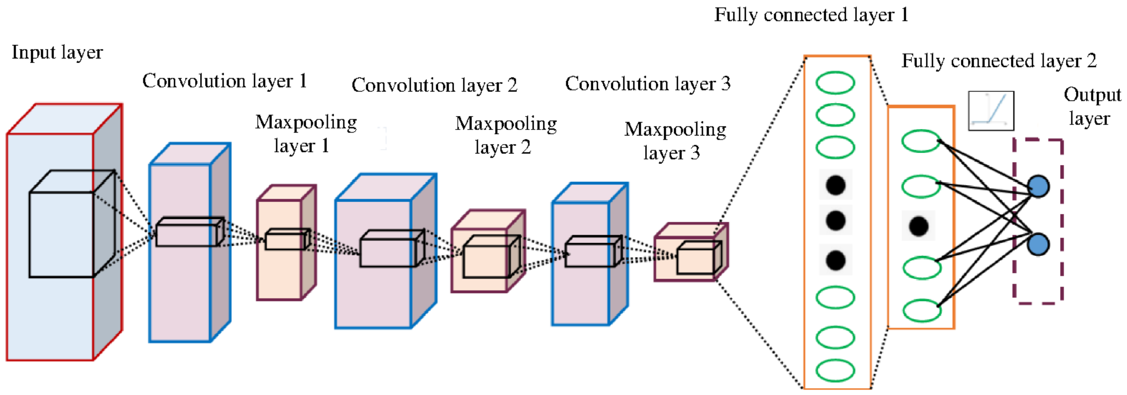
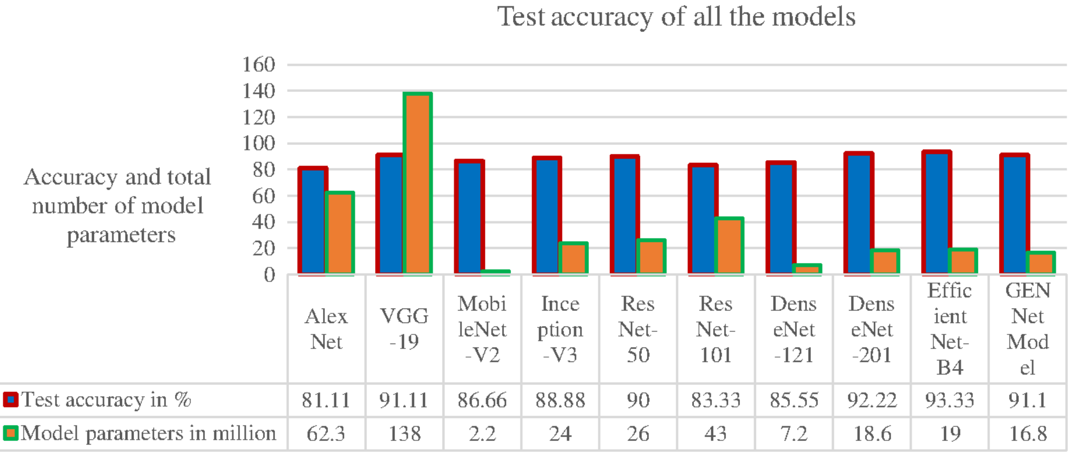


Figure 31: Structural layer-wise architecture of newly proposed GENNet architecture.

- The overall test accuracy along with the number of model parameters of all the architectures is shown in Figure 32. The EfficientNet-B4 model achieved the highest test accuracy of 93.3% followed by the DenseNet-201 and the proposed GENNet network which has achieved an overall testing accuracy of 92.2 and 91.1% however, GENNet architecture is good for a compute-constrained thermal gender classification use-case as it performs significantly better than other low-parameter models.



	AlexNet	VGG-19	MobileNet-V2	Inception-V3	ResNet-50	ResNet-101	DenseNet-121	DenseNet-201	EfficientNet-B4	GENNet Model
🟦 Test accuracy in %	81.11	91.11	86.66	88.88	90	83.33	85.55	92.22	93.33	91.1
🟧 Model parameters in million	62.3	138	2.2	24	26	43	7.2	18.6	19	16.8

Figure 32: Validation accuracy and model parameters of all the CNN architectures.

The complete working methodology and experimental details of this work is published in the Journal of Electronic Imaging (JEI) by SPIE titled “Performance Estimation of the State-of-the-Art Convolution Neural Networks (CNN) for Thermal Images-Based Gender Classification System” [11]. This paper presents and summarizes the training and validation results of all the models along with the newly proposed GENNet architecture. The copy of the published paper based on this section is presented in Appendix C.

Chapter 5

Contribution To Development of Out-cabin Thermal Object Detection/ Classification System using SoA YOLO Framework

Object detection in thermal infrared spectrum provides a more reliable data source in low-lighting conditions and different weather conditions, as it is useful both in-cabin and outside for pedestrian, animal, and vehicular detection as well as for detecting street signs & lighting poles. The core contribution in this work includes the design and development of smart thermal perception system for the automotive sensor suite by exploring and modifying state-of-the-art object detection and classifier framework on thermal vision with various distinct classes for advanced driver-assistance systems (ADAS). This work is carried under WP-8 of the Heliaus project focused towards out of cabin applications.

5.1. Research Objectives

Advanced Driver-Assistance Systems (ADAS) has become a developing consumer technology product and the evolution of this technology over time intends to provide extended safety advantages and trustworthy means of transportation. Numerous technologies are directly associated with ADAS which includes, sensor fusion for real-time data logging, and object/ obstacle detection and tracking system using advanced machine learning algorithms. This will enable the drivers to monitor the external environment, sense external objects, and predict results that the driver needs to be aware of thus providing a deeper perception of the entire road network. In this work, we have particularly focused on developing an out-cabin thermal object detection/ classification system as forward sensing (F-sense) system for vehicular technology. The main advantages of such type thermal environmental perceptions include.

1. The system can sense and analyze its environment reliably and accurately
2. Further such systems can interact with the driver to request him to intervene appropriately

Current ADAS largely rely on computer vision and machine learning which uses visible (RGB) or RGB + near-infrared (NIR) cameras as a sensor. The alternative is ultrasonic, lidar, and radar-based [70] hardware sensors. Practical systems often leverage both camera + radar and lidar. However, the mentioned sensors and imaging modalities have some of their limitations. For instance, lidar provides a sparse three-dimensional (3D) map of the environment, but small objects like pedestrians and cyclists are difficult to detect especially when they are at a distance [71]. The RGB camera operates inadequately in unfavorable illumination conditions such as low lighting, sun glare, and glare from the headlight beam. Radar has a low spatial resolution to detect pedestrians accurately [71]. Also, large objects such as cars can saturate the performance of the receiver if they are closer to the transmitter, and lastly, the performance of the radar is severely affected as the radio signals can face enough natural interference.

The current advancement in bolometer technology has led to cheaper yet more effective solutions in the form of the development of uncooled thermal cameras. These cameras can replace the conventional sensors such as Lidar and Radar or can be integrated with the existing hardware sensors to provide more comprehensive information about road-side environmental perception. Since thermal imaging is invariant to illumination changes, occlusions, and shadows it provides improved situational awareness that results in deploying more robust, reliable, and safe systems for intelligent vehicular systems. AI-based imaging pipelines are commonly used for designing intelligent objection detection-based video perceptions systems. To accomplish this goal we had focused on using a state-of-the-art YOLO-v5 framework [72] for training and deploying thermal object detection system on GPU and Edge-GPU platforms. This work belongs to work package 4 (WP-8) of the Heliaus project [1] which is focused on the development and validation methodologies of the thermal-IR system for out-cabin vehicular applications.

The main reason for selecting the YOLO-v5 framework for thermal data as compared to all the previous versions of YOLO released is that YOLO-v5 is different, as this is a PyTorch implementation rather than a fork from the original Darknet library. Moreover, the YOLO v5 has a Cross-Stage-Partial (CSP) backbone and PA-NET neck. The foremost improvements include mosaic data augmentation and auto-learning bounding box anchors. The detailed comparative analysis of the recently released Yolo-v5 with all the previous versions is presented in Table 7.

Table 7: Comparison Analysis of Previous Yolo versions with Yolo-v5

Yolo Version	Training Dataset	Validation mAP	FPS	Implementation Framework
YOLO [73]	VOC 2007 + 2012	63.4	45	DarkNet
YOLO-v2 (608x608) [74]	MSCOCO	48.1	40	DarkNet
YOLO-v3 (608x608) [74]	MS COCO	57.9	20	DarkNet
YOLO-v4 (608x608) [75]	MSCOCO	43.5	62	DarkNet
YOLO-v5 (640x640) [72] X Large Model	MSCOCO	68.9	83	PyTorch

It can be observed from the above table that YOLO-v5 has comparatively achieved better validation results in terms of the highest mean average precision and frames per second on the COCO dataset as compared to the previous version of the YOLO framework.

5.2. Summary and Discussions of Contributions

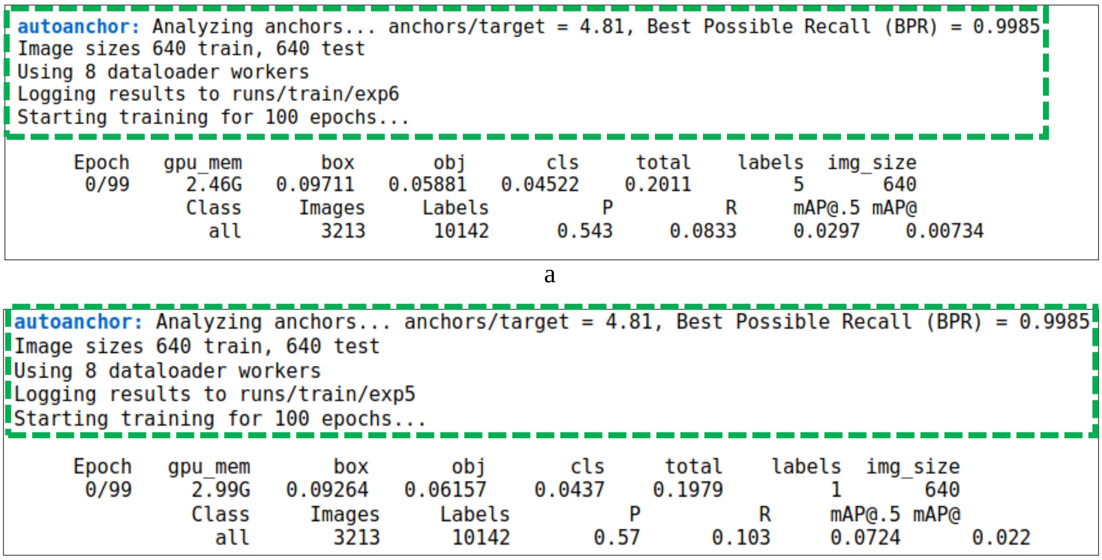
Object detection algorithms are normally trained on one or two datasets for various computer vision applications. However, such type of method is not successful when coming to ADAS application. This is due to the fact that including image data from one specific dataset that is collected in certain areas, means that an object detector will be trained or fine-tuned on mentioned datasets which may not perform optimally when it is tested with another dataset that contains image data gathered from another city/ environment with different scene contexts. To cater to this challenge, we have selected five different thermal datasets for optimal training of various network variants of the YOLO framework. These datasets include OSU-thermal [76], CVC-09 [77], KAIST multispectral dataset [78], FLIR-ADAS dataset [79], and locally acquired object detection data as discussed in chapter 3 section 3.2.2 [17]. The foremost contributions of this work are highlighted in bullet form.

- In the first phase, we have performed the optimal training of four different network variants of Yolo-v5 frameworks. These models include small variant, medium variant, large variant, and extra-large variant. The individual model attributes of five different network variant available in the YOLO-v5 framework [72] is provided in Table 8.

Table 8: Yolo-v5 Network Variants Attributes

S. No	Model	Image size (pixels)	Parameters (million)	Flops @640 (B)	mAP @COCO Dataset (%)
1	Small	640	7.2	16.5	56.0
2	Medium	640	21.2	21.2	63.9
3	Large	640	46.5	46.5	67.2
4	X-Large	640	86.7	86.7	68.9

- As mentioned in section 5.1 the YOLO-v5 framework is published with notable improvements which include auto-learning bounding box anchors and mosaic data augmentation. This means that rather than computing the anchor values manually as was required in the previous versions of YOLO for optimal training of the CNN networks the anchors are computed automatically before the training process. The anchors are evaluated against the training dataset in combination with the training settings which incorporates image size (640X640 in our case), and the number of classes (Nc=6 in our case). If the Best Possible Recall (BPR) is below the threshold then the anchors are determined not to be a good fit for custom training data, and thus new anchors are computed to replace them using a genetic algorithm optimizer initialized by K-means centroid-based algorithm. This is all transparent and fully displayed at the beginning of the training workspace as shown in Figure 33.



```
autoanchor: Analyzing anchors... anchors/target = 4.81, Best Possible Recall (BPR) = 0.9985
Image sizes 640 train, 640 test
Using 8 dataloader workers
Logging results to runs/train/exp6
Starting training for 100 epochs...

     Epoch   gpu_mem       box       obj       cls     total    labels  img_size
      0/99     2.46G   0.09711   0.05881   0.04522    0.2011         5       640
               Class    Images    Labels         P         R   mAP@.5 mAP@
                 all      3213     10142     0.543    0.0833    0.0297    0.00734
```

a

```
autoanchor: Analyzing anchors... anchors/target = 4.81, Best Possible Recall (BPR) = 0.9985
Image sizes 640 train, 640 test
Using 8 dataloader workers
Logging results to runs/train/exp5
Starting training for 100 epochs...

     Epoch   gpu_mem       box       obj       cls     total    labels  img_size
      0/99     2.99G   0.09264   0.06157    0.0437    0.1979         1       640
               Class    Images    Labels         P         R   mAP@.5 mAP@
                 all      3213     10142      0.57     0.103    0.0724     0.022
```

b

Figure 33: Auto-learning bounding box anchors feature in YOLO-v5 framework, a) auto anchor processing during the training process of the small model with BPR of 0.985, b) auto anchor processing during the training process of the medium model with anchors = 4.81 and BPR of 0.9985.

- Secondly, for optimum convergence/ adaptation of CNN models on thermal data besides the auto-learning bounding box anchors, we have used two different optimizers which include SGD and Adam as discussed in section 3A of our published paper [18] (Appendix E). After doing the initial experimental training process we have shortlisted SGD optimizer for the training process of all the network variants as it performs significantly better when compared to Adam optimizer. The graphical results comparisons of small and large network variants extracted via Tensorboard are demonstrated in Figure 34. It can be observed from the below figure that in the case of both the models the training mean average precision curve is much higher using the SGD optimizer.

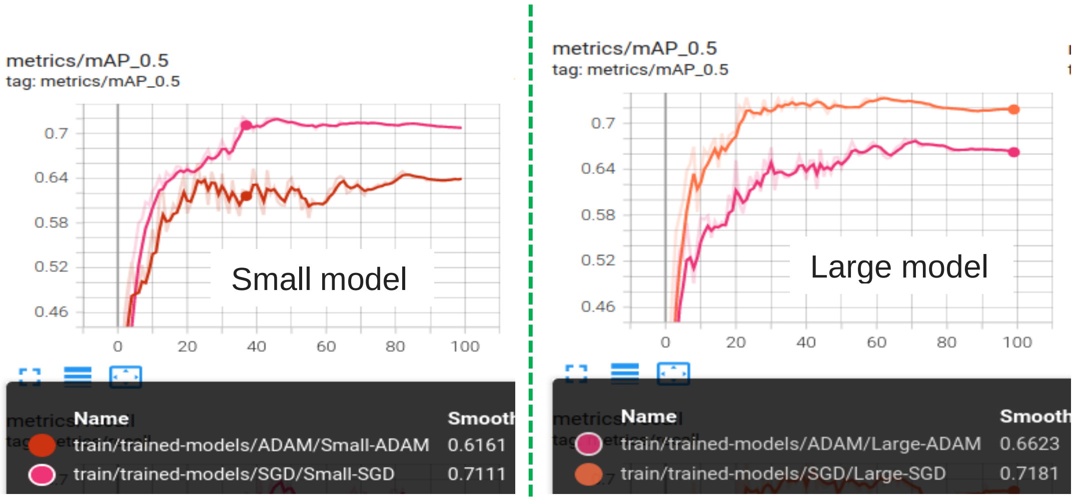


Figure 34: Small and large model training results comparison using SGD and ADAM optimizer.

- During the training phase rather than using fixed learning rate we have used one cyclic learning rate by defining base (lower bound) and maximum learning rate (upper bound) values. It works by updating the LR value back and forth between the defined bound values after every batch. Moreover, during the training, we have employed different data transformation methods to bring enough data diversity/ variation which will eventually be helpful for the deep learning models to learn thermal data features more robustly. Alongside conventional data transformational methods, we have used the Mosaic Augmentation method. It is an advanced form of image augmentation operation which works by combining different training samples in one image with varying ratios. It helps the network to learn how to identify the objects at a smaller scale than normal.

- In the second phase, the performance evaluation of all the trained models is validated using three different test approaches which include test-time with no augmentation (TTNA), test-time augmentation (TTA), and test-time with model ensembling (ME). Model ensembling or ensembling engine refers to using multiple trained networks in a parallel manner to produce one optimal predictive inference model (Appendix D). In this study, we have tested the performance of individually trained variants of the Yolo-V5 framework and selected the best combination of models which in turn helps in achieving better mean-average precision (mAP) scores for the validation purpose on test data. Figure 35 shows the structural block diagram of the proposed ensembling inference engine for the thermal object detection/ classification system.

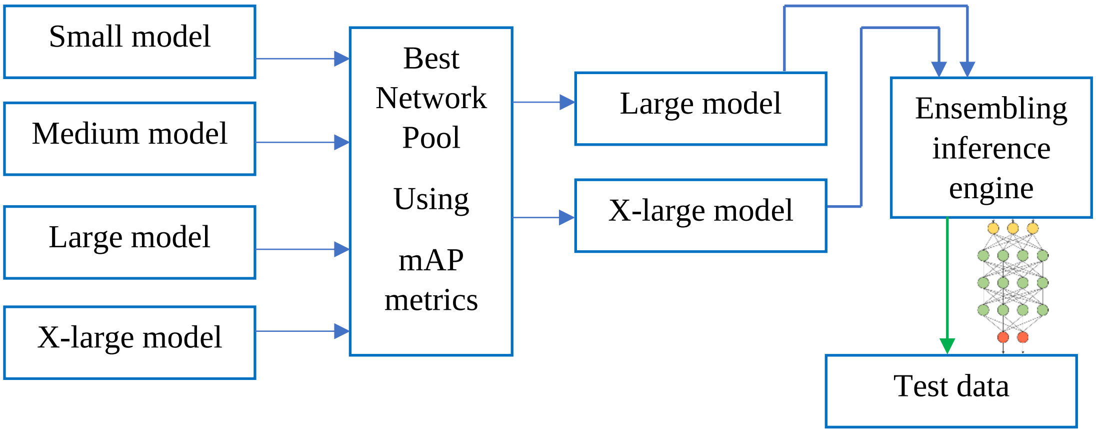


Figure 35: Structural block diagram representation of ensembling inference engine based on the combination block of large and x-large network variants.

Figure 36 shows the inference results on nine different challenging thermal frames with complex scenarios like multiple objects with overlapping classes, object scale and viewpoint variations, and different weather conditions. These frames are selected from public test data as well as locally acquired data as discussed in chapter 3 section 3.3.2. It can be observed that the ME engine has performed significantly well on complex thermal frames.

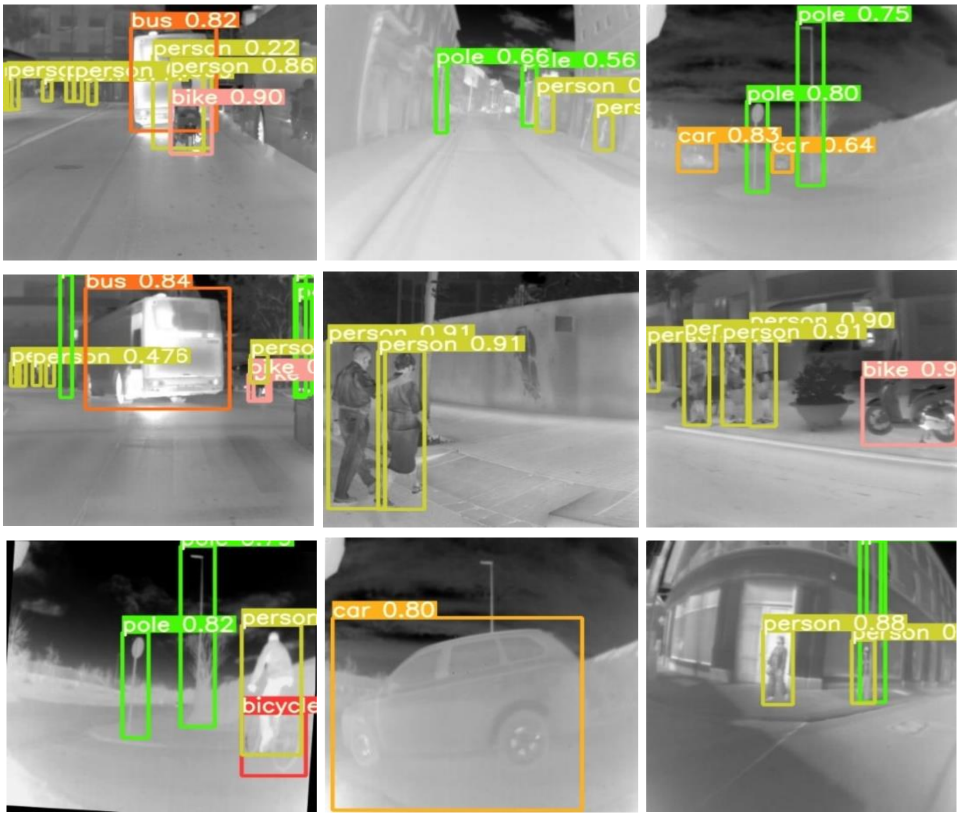


Figure 36: Inference results on nine different thermal frames using the model ensembling inference engine.

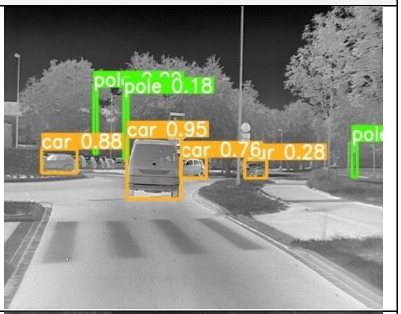
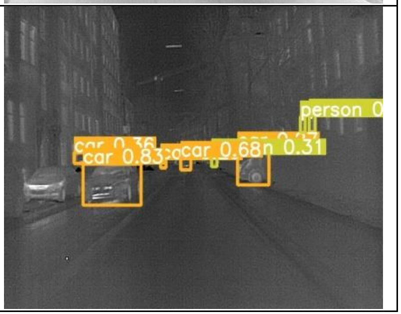
- For a rigorous validation test of all the thermally tuned network variants of the Yolov5 framework, four different quantitative metrics have been employed. These include mean average precision, overall model precision, recall, and inference time required per frame. The experimental validation details can be found in section IV of Appendix D.

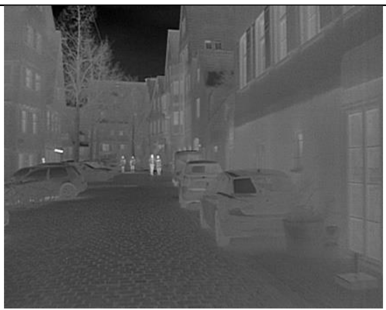
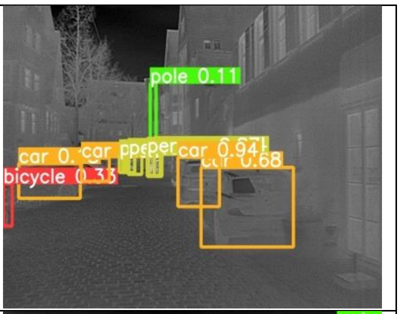
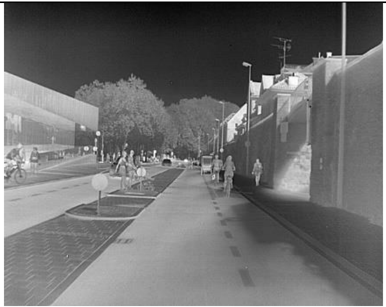
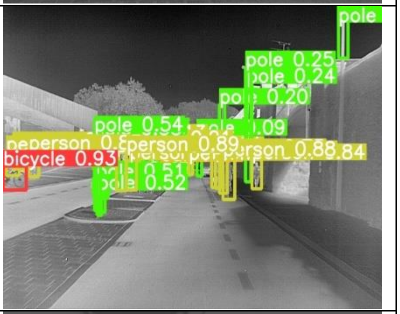
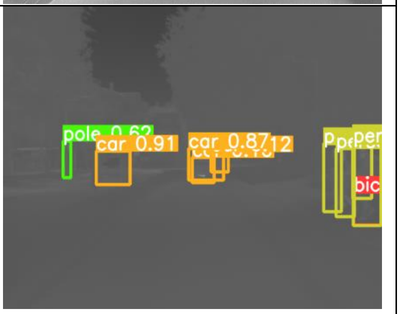
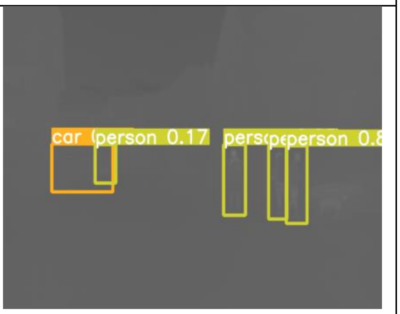
The complete training methodology along with the detailed experimental results is presented and published in IEEE Access Journal titled "Object Detection in Thermal Spectrum for Advanced Driver-Assistance Systems (ADAS)". The copy of the published paper based on this section is attached and presented in Appendix E of this report.

5.3. *Further Experimental Results on Denso Out-Cabin Thermal Data*

In addition to our work discussed in section 5.2, the efficacy of thermally tuned object detection models was also validated on extensive out-cabin thermal data acquired by Denso Germany. Denso was working as an industry consortium partner in the Heliaus project. DENSO [80] expertises in camera base algorithm development and collecting new diversified thermal datasets for algorithms training, and validation. The extensive out-cabin data gathering and thus the creation of relevant and diversified thermal image sets was also a valuable outcome of the project. Table 9 shows the inference results on various challenging thermal frames acquired by Denso in different environmental and weather conditions using the prototype 640x480 thermal camera developed by Lynred France under the Heliaus project.

Table 9: Inference Results on Denso Out-Cabin Thermal Data

Frame and Algorithms Details	Input image	Inference results
Frame details: Acquired in the daytime with shutterless camera calibration and applying image correction pipelines Models used: X-large model using test-time with no augmentation		 pole 0.18 car 0.88 car 0.95 car 0.76 r 0.28 pol
Frame details: Acquired in the nighttime with shutterless camera calibration and applying image correction pipelines Models used: X-large model using test-time with no augmentation		

Frame details: Acquired in the nighttime with shutterless camera calibration and applying image correction pipelines Models used: X-large model using test-time augmentation		 pole 0.11 car 0. car car 0.94 car 0.68 bicycle 0.33
Frame details: Acquired in the daytime with shutterless camera calibration and applying image correction pipelines Models used: large model using test-time augmentation		 pole pole 0.25 pole 0.24 pole 0.20 pole 0.54 1.09 person 0.89 person 0.88.84 bicycle 0.93 pole 0.51 pole 0.52
Frame details: Acquired in the daytime with shutterless camera calibration but without applying image correction pipelines Models used: small and x-large model using model ensembling approach		
Frame details: Acquired in the nighttime with shutterless camera calibration but without applying image correction pipelines Models used: small and x-large model using the model ensembling approach		

It can be observed from Table 9 that trained networks performed well on some of the challenging and newly acquired (unseen) thermal frames. The trained detectors show robust results on thermal frames without applying the image correction pipeline as shown in row 5 and row 6 of Table 9 thus detecting and predicting most of the class objects with high confidence scores values.

Chapter 6

Contribution To Deployment of Thermal Object Detection/ Classification Framework on GPU & EDGE-GPU devices using Advanced Neural Optimization Methods

This chapter will highlight the main contributions by further optimizing and deploying the neural networks on GPU as well as resource-constrained edge devices which include Nvidia Jetson Nano and Nvidia Jetson Xavier development boards. The thermally tuned YOLO architectures are further optimized using SoA inference accelerator to produce higher frames per second (FPS) and lower inference time. The main reason for performing the quantization process is to check the feasibility of thermally tuned object detection models for real-time onboard testing when deploying the models on edge hardware for the automotive sensor suite. This work is carried out under WP-4 of the Heliaus project focused towards the deployment of neural network-based processing frameworks on dedicated embedded devices.

6.1. Research Objectives

Object detection algorithms have currently encountered numerous challenges due to the demands of high inference speed and accuracy, especially for real-time scenarios. While the inference speed depends mainly on hardware resources such as GPU and Edge-GPU devices and the complexity of the network. On the other hand, the accuracy depends mainly on the algorithm adopted for the system. In the current scenario where the hardware technology is evolving and developing rapidly, the object detection algorithms are expected to perform at higher speeds in near future. Currently, the best preference for deploying the object detector algorithms with the automotive sensor suite is the selection of a network that can balance efficiently the inference speed and detection accuracy. To achieve this purpose, we selected various network variants of state-of-the-art YOLO-v5 [72] as shown in Table 8 (in chapter 5) for the thermal object detection task. The performance of successfully converged models on thermal data is validated on different hardware resources which include GPU, Nvidia Jetson Nano [52], and Nvidia Xavier [53] edge development boards for real-time onboard feasibility evaluations.

6.2. Summary and Discussions of Contributions

In this work, the core emphasis is to achieve good speed (reduced inference time) since we are typically focusing to run the networks on embedded architectures as well as adequate accuracy level (mAP) with the reduced false alarm rate. To accomplish this task and successfully deploy the thermal-YOLO architectures on GPU & edge embedded architectures we have adapted SoA techniques for performing model optimization without compromising/ affecting the overall model accuracy. The main contributions of this work are highlighted in bullet form.

- In the first phase, we have performed the precise quantitative test of all the thermally tuned models by analyzing the performance of these architectures using four different metrics which includes include recall, precision, mean average precision (mAP), and frames per second rate (FPS). For this purpose three different confidence thresholds intervals i.e 0.4, 0.2, and 0.1, and the intersection of union (IoU) intervals i.e 0.6, 0.4, & 0.2 were employed. It is noticeable from Table 10 that by decreasing the confidence threshold value of the thermally tuned small-YOLO variant the mAP increases gradually. However, by decreasing the confidence threshold too much, we can lead to high mAP but with high false alarms rate (wrong bounding box locations) and higher inference speed which is eventually not good for an optimal thermally tuned model focused to be deployed on embedded architectures.

  Table 10: mAP of the small model using different confidence thresholds

Network	Confidence Threshold	mAP
Small variant	0.4	45%
Small variant	0.2	47.4
Small variant	0.1	48.3

  The complete results of all the four quantitative metrics for all the modes are presented in section V-C of Appendix F.

- In the second phase, model optimization is performed using the SoA TensorRT inference accelerator to implement a high-speed inference network on SoA embedded GPU boards (Jetson and Xavier) with evaluations. The primary motivation for this is to boost the FPS rate for real-time evaluations and on-board feasibility testing on edge devices. Second, it uses several optimization strategies to reduce onboard memory footprints on the target device. Figure 37 shows the block diagram representation of the proposed optimized inference engine for embedded architectures. Whereas Figure 38 shows the inference results on various thermal frames using the TensorRT inference accelerator engine [51]. The comprehensive experimental details of model optimization is presented in section VI of Appendix F.

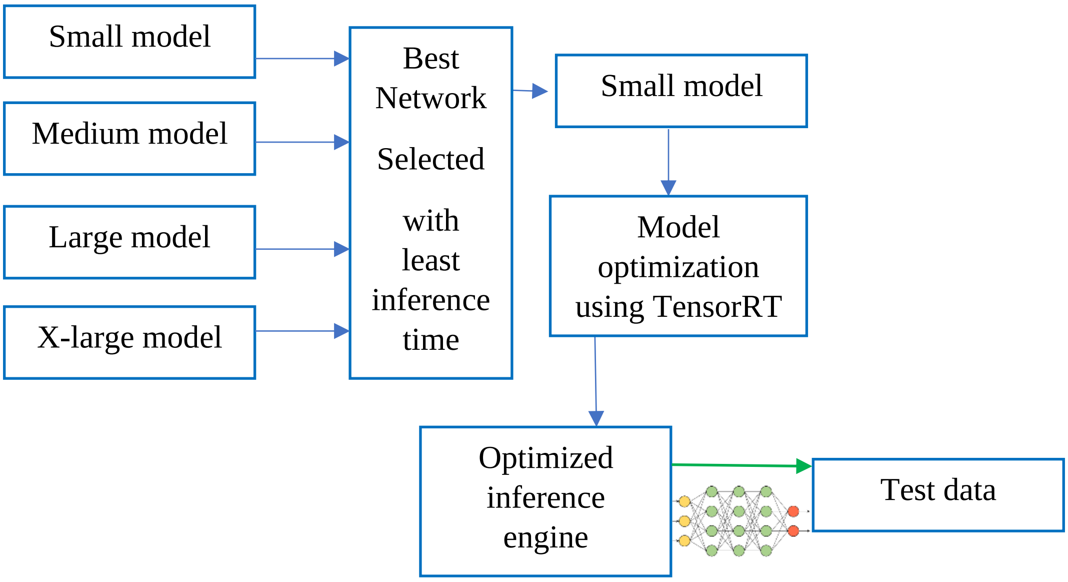

Small model
Medium model
Large model
X-large model
Best Network Selected with least inference time
Small model
Model optimization using TensorRT
Optimized inference engine
Test data

Figure 37: Structural block diagram representation of TensorRT based optimized inference engine using the small network variant for deployment on edge architectures.

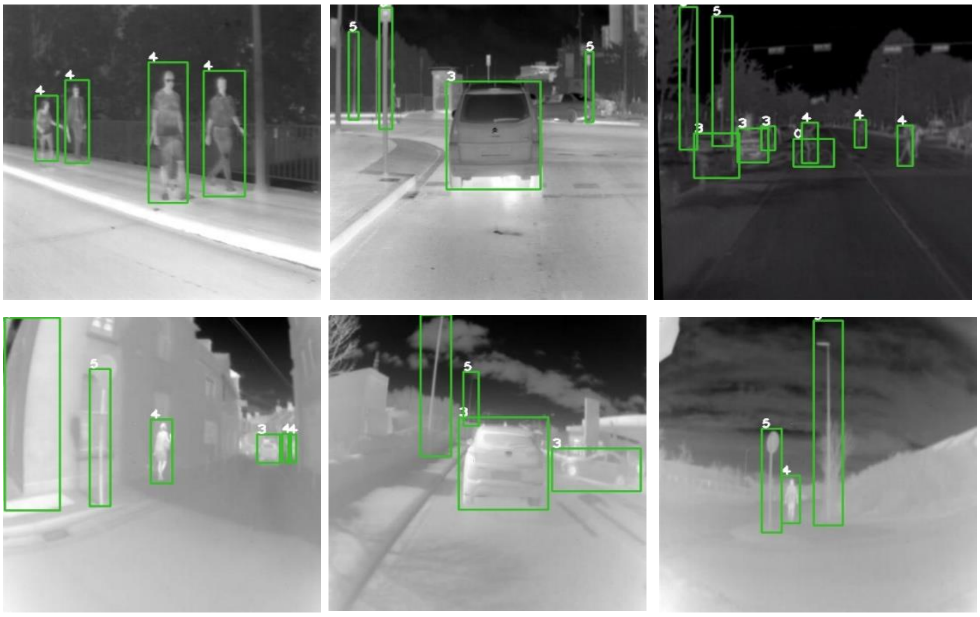


Figure 38: Inference results on six different thermal frames using the TensorRT optimized inference engine on Nvidia Jetson Nano and Nvidia Jetson Xavier boards.

The optimized version of the smaller network variant achieved 60 FPS on the Nvidia Jetson Xavier development board and 11 FPS on the Nvidia Jeston Nano board.

- Moreover while running the inference and quantitative test we closely monitor the temperature ratings of different hardware peripherals on both Edge-GPU platforms using a specialized jetson-stats open-source python library. It is done to avoid the overheating effect which can harm the onboard processor and affect the overall operational capability of the system.

The complete working methodology along with the detailed experimental results is presented and published in IEEE Transactions on Intelligent Vehicles Journal titled “Evaluation of Thermal Imaging on Embedded GPU Platforms for Application in Vehicular Assistance Systems” [19]. The copy of the published paper based on this section is attached and presented in Appendix F of this report.

Chapter 7

Additional Contributions

This chapter will summarize my additional contributions along with some of my secondary publications related to these contributions.

7.1. Discussions of Contributions towards Human Thermography

In the first part, I explored infrared thermal imaging for human thermography and cancer diagnosis using thermal and RGB Dermoscopic imaging. Human thermography is an integral medical diagnostic tool for detecting heat patterns and measuring quantitative temperature data of the human body as shown in Figure 1 of chapter 2. It can be used in conjunction with other medical diagnostic procedures for getting comprehensive medication results. Infrared Thermography (IRT) plays a vital role in detecting abnormal temperature patterns in human organs which can be further used for detecting low-risk as well as fatal diseases in their early stages. Human thermography can be effectively used for a wide range of disease detection and classification, in both male and female gender some of which are discussed in Appendix G.

Alongside the advantages of human thermography, there are also certain disadvantages of thermography. The advantages and disadvantages of human thermography are discussed below.

Advantages

- It is a type of non-invasive and painless technique.
- It provides a user-friendly and easy seating examination process for the patients.
- The minimal time required to perform the overall test is about 2 to 3 min.
- The test results are much easier to judge by monitoring the difference in colour changes (gradient: −0.05 °C)
- There are many different methods available to store the thermographic results also known as thermograms. Such as we can use simply use paper printing, Xerox paper printing, or coated with a material that changes colour on heating also known as thermal printing. Secondly, we store the thermograms on modern storage techniques such as magnetics devices which include hard-drive, compact discs, and flash drives.
- One of the biggest advantages of thermography is that it does not produce any harmful radiations like conventional medical examinations such as Computed Tomography (CT) scan machines, mammography, and X-Ray test which uses low-dose X-rays to take pictures from inside the human organ.

Disadvantages

- The hardware required for performing thermographic tests such as high resolution cooled thermal cameras are very expensive
- The sensitivity and resolution of the camera reduces with variation in distance and angle of view
- There are many different factors such as environmental and body factors which can affect the overall thermal imaging results.

The core importance of IRT is highlighted in one of our publications titled “Infrared Imaging for Human Thermography and Breast Tumor Classification using Thermal Images” published in the 31st Irish Signals and Systems Conference (ISSC) [20]. In the first portion of this publication, we have discussed various disease diagnoses using thermography whereas the second part of the paper discusses breast tumor classification system using computer vision and machine learning-based algorithms. The copy of the published paper based on this section is attached and presented in Appendix G of this report.

In addition to that, I have also worked on skin cancer diagnosis using RGB dermoscopic images. Dermoscopic diagnosis refers to a non-invasive skin imaging method, which has become an essential tool in the diagnosis of melanoma and other pigmented skin lesions. However, performing dermoscopy using conventional methods may reduce the diagnostic accuracy which can lead to more chances of errors. These errors are generally caused by the complexity of lesion structures and the subjectivity of visual interpretations. In this work, I have proposed AI-based computer-aided diagnosis system for skin cancer classification in prodromal stages [21]. The copy of the published paper based on this section is attached and presented in Appendix H of this report.

7.2. Discussions of Contributions towards Monocular Depth Estimation

Depth estimate is an important step in inferring scene geometry from two-dimensional images. Given only a single RGB frame as input, the objective of monocular depth estimation is to estimate the depth value of each pixel or infer depth information. In this work, we have investigated the facial depth datasets and loss functions generated in the field of computer vision for facial depth estimation problems. In the next step, we have presented the implementation details of how neural depth networks work, as well as their associated evaluation matrices, which are summarized in our published study [22]. In addition to this, a SoA neural architecture for facial depth estimation is proposed, along with a comparative evaluation with other SoA methods [22]. The complete study along with experimental details titled ‘Towards Monocular Neural Facial Depth Estimation: Past, Present, and Future’ can be found in Appendix I of this report.

Chapter 8

Conclusion and Future Work

In this thesis, we addressed the problem of designing a smart thermal perception system for driver assistance applications thus providing enhanced safety and reliability features using a low-cost yet reliable LWIR uncooled thermal camera module. Such type of system can be beneficial and integrated with the existing automotive sensor suite for advanced vehicular systems.

8.1. Experimental Outcomes of this Thesis

In earlier chapters, we have introduced various works undertaken during the course of this Ph.D. research. Here we summarise the main findings of this work, placed in the context of the main goals & objectives presented in the introduction chapter (chapter 1) of this report. The core experimental outcomes of this work are as follows.

1. In chapter 3 we have presented our contributions for large scale thermal synthetic data generated using the composite structure of advanced computer vision algorithms which includes data transformation/ augmentation, synthetic data generation using styleGAN based generative adverisal network, and single 2D image to 3D thermal face reconstruction using end to end Position Map Regression Network (PRNet).

2. In chapter 4 we have presented our contributions towards the design and validation of driver/ occupant gender classification system in the thermal spectrum by retraining and fine-tuning a set of SoA pretrained deep convolutional neural networks (CNN) which includes Alexnet, VGG, MobileNet, ResNet, Inception-v3, DenseNet, and EfficientNet architectures. In addition to that, we have proposed an entirely new CNN (GENNet) designed for the thermal gender classification task, and its performance was benchmarked with other SoA pretrained architectures. The EfficientNet model attained the maximum testing accuracy of 93% followed by the DenseNet-201 and the proposed GENNet network which had achieved an overall testing accuracy of 92% and 91% respectively. However, the performance evaluation of GENNet architecture shows that a smaller, more lightweight network can perform better when trained directly on data for a specific imaging application.

3. In chapter 5 we have presented our contributions towards adaptation and validation of a state-of-the-art object detection/ classification framework (Yolo-v5) for designing smart thermal perception system with seven distinct classes including stationary as well as moving objects. Four different network variants were trained on four different public datasets as well as locally gathered novel test data. Moreover, three different testing approaches were used for rigorous validation of all the trained networks which includes test-time with no augmentation (TTNA), test-time augmentation (TTA), and test-time with model ensembling (TTME) methodology. Lastly, a new model ensemble-based inference engine is proposed using the combination of X-large and large model which proves to be the best network coupler thus producing the results as one optimal inference engine. The TTME inference engine further improves the accuracy metrics on overall test data. Also, the performance of trained detectors shows robust detection

and classification results when tested on large-scale out-cabin thermal data acquired by Denso. This shows that trained models have learned optimal feature information to perform precisely on a wide range of test data.

4. In chapter 6 our contributions mainly focus on further optimization of thermally tuned object detection models. This is done since we are typically focusing to run the networks on low-power embedded architectures for real-time onboard feasibility testings. The process of model optimization is attained using TensorRT inference accelerator to implement a fast inference compute engine on embedded GPU boards which included Nvidia Jetson Nano and Nvidia Xavier development boards. After the successful execution of model optimization, we were able to achieve a frame rate of 11 FPS on (4 core CPU) Nvidia Jetson and subsequently a very high frame rate of 60 FPS on 6 core CPU) on the Nvidia Jetson Xavier board. The current SoA GPU engines are adequate for running optimized machine learning models but, realistically, they are running very hot and further improvements are needed before real-time, energy-efficient AI can be implemented in automotive use cases.

It is important to mention that during the entire experimental training process of a wide range of CNN models we had particularly focused on two different SoA optimizers which include SGD [81] and ADAM [82]. The best optimizer was selected in accordance with other network hyperparameters and training data tasks. Both of these optimizers have their benefits. SGD optimizers mostly operate in a small-batch regime wherein a fraction of the training data, usually, 32-512 data points, is sampled to compute an approximation to the gradient. On the other hand, the Adam optimizer combines the best properties of the AdaGrad and RMSProp algorithms to provide an optimal algorithm that can handle sparse gradients on noisy data.

8.2. Dataset Contributions

The dataset contributions of this work resulted in the form of novel indoor thermal facial data and out-cabin object detection data collection as discussed in chapter 3. The summarized details of these datasets are provided below.

1. Indoor data is comprised of thermal facial images acquired from both male and female subjects who agreed to take part in this research work. The data was recorded in indoor environmental conditions with controlled lighting conditions. This data is useful to carry out empirical work in relation to designing effective in-cabin driver and passenger monitoring systems for enabling enhanced safety features. The collected dataset was used in various types of experimental work which included thermal gender classification, thermal face detection, and eye detection for generating drowsiness alerts using computer vision-based algorithms.

2. A novel out-cabin thermal data is acquired and annotated consisting of >35k distinct 640x480 frames using a prototype thermal camera module based on micro-bolometer technology. The overall data is recorded in two different methods (M1 & M2) such that by mounting the camera at a fixed place and in the second method the data is recorded by mounting the camera on an electric car. The recorded thermal frames incorporate data variations in the form of diverse weather and environmental conditions.

8.3. Future Work

Many different adaptation tasks related to 3D depth estimation using thermal images, training and validation testings of mask segmentation and instance segmentation algorithms on thermal data, and related experimental work have been left for the future due to lack of time (i.e. the experiments with real + synthetic data are usually very time consuming, requiring even days to optimally retrain a model from scratch on thermal data). Future work concerns deeper analysis for optimal training and validation mechanisms of deep learning architectures, new proposals to try different methods, or simply curiosity. As the possible future directions, these are some of the ideas that I would like to further explore and give the opportunity to research the community to further work on these aspects which are as follows.

1. Design and implementation of SoA mechanism to build high-quality 3D depth maps from multiple thermal images using depth estimation models which can be beneficial for in-cabin and out-cabin applications.

2. Extensive ADAS datasets should be gathered and open-sourced with synchronization of various imaging modalities such as depth data, event data, thermal data, near-infrared data. These data can be beneficial for optimal training of deep learning models for in-cabin and out-cabin applications.

3. Moreover, the joint fusion of RGB and thermal imaging can be advantageous and taken into consideration for better training of deep learning networks for in-cabin as well as out-cabin applications. More experiments need to be made to come up with an optimal solution for producing fused images thus building a single multiset or multiway structure with all images involved or connecting the related individual images through regression and deep learning models.

4. The deployment of optimized trained networks on more powerful single-board edge devices with a higher flop rate and less operating power for optimal performance, thus tailoring it for real-time onboard deployments.

5. Finally, the current out-cabin work just focuses on object detection/ recognition, but it can be modified and enhanced to incorporate image segmentation, road lane detection, traffic signal, and road signs classification, and object tracking for estimating objects distance from vehicle thus providing comprehensive data for enhanced driver assistance.

References

[1]. Heliaus European Union Project, [Online] Available: https://www.heliaus.eu/ (Last accessed on 8th March 2022).

[2]. Besbes, Bassem, et al. "Pedestrian detection in far-infrared daytime images using a hierarchical codebook of SURF." *Sensors* 15.4 (2015): 8570-8594.

[3]. Ziebinski, Adam, et al. "A survey of ADAS technologies for the future perspective of sensor fusion." *International Conference on Computational Collective Intelligence*. Springer, Cham, 2016.

[4]. Tompson, Jonathan, et al. "Efficient object localization using convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[5]. Chauhan, Nitin Kumar, and Krishna Singh. "A review on conventional machine learning vs deep learning." *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE, 2018.

[6]. Pranav, K. B., and J. Manikandan. "Design and Evaluation of a Real-Time Face Recognition System using Convolutional Neural Networks." *Procedia Computer Science* 171 (2020): 1651-1659.

[7]. Židek, Kamil, et al. "Recognition of assembly parts by convolutional neural networks." *Advances in Manufacturing Engineering and Materials*. Springer, Cham, 2019. 281-289.

[8]. Van der Meulen, Rob. "What edge computing means for infrastructure and operations leaders." Smarter with Gartner (2018), [Online] Available: https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders (Last accessed on 8th March 2022).

[9]. Farooq, Muhammad Ali, and Peter Corcoran. "Proof-of-Concept Techniques for Generating Synthetic Thermal Facial Data for Training of Deep Learning Models." *2021 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2021.

[10]. Farooq, Muhammad Ali, and Peter Corcoran. "Generating Thermal Image Data Samples using 3D Facial Modelling Techniques and Deep Learning Methodologies." *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020.

[11]. Farooq, Muhammad Ali, Hossein Javidnia, and Peter Corcoran. "Performance estimation of the state-of-the-art convolution neural networks for thermal images-based gender classification system." *Journal of Electronic Imaging* 29.6 (2020): 063004.

[12]. K. Panetta et al., “The tufts face database,” [Online] Available: http://tdface.ece.tufts.edu/ (Last accessed on 29 January 2022).

[13]. Panetta, Karen, et al. "A comprehensive database for benchmarking imaging systems." *IEEE transactions on pattern analysis and machine intelligence* 42.3 (2018): 509-520.

[14]. KM, Shreyas Kamath, et al. "TERNet: A deep learning approach for thermal face emotion recognition." *Mobile Multimedia/Image Processing, Security, and Applications 2019*. Vol. 10993. International Society for Optics and Photonics, 2019.

[15]. Espinosa-Duró, Virginia, et al. "A criterion for analysis of different sensor combinations with an application to face biometrics." *Cognitive Computation* 2.3 (2010): 135-141.

[16]. Espinosa-Duró, Virginia, Marcos Faundez-Zanuy, and Jiří Mekyska. "A new face database simultaneously acquired in visible, near-infrared and thermal spectrums." *Cognitive Computation* 5.1 (2013): 119-135.

[17]. Muhammad Ali Farooq, Waseem Shariff, Faisal Khan, Peter Corcoran, Cosmin Rotariu, March 26, 2022, "C3I Thermal Automotive Dataset", IEEE Dataport, doi: https://dx.doi.org/10.21227/jf21-rt22.

[18]. Farooq, Muhammad Ali, et al. "Object Detection in Thermal Spectrum for Advanced Driver-Assistance Systems (ADAS)." *IEEE Access* 9 (2021): 156465-156481.

[19]. M. A. Farooq, W. Shariff and P. Corcoran, "Evaluation of Thermal Imaging on Embedded GPU Platforms for Application in Vehicular Assistance Systems," in *IEEE Transactions on Intelligent Vehicles*, doi: 10.1109/TIV.2022.3158094.

[20]. Farooq, Muhammad Ali, and Peter Corcoran. "Infrared imaging for human thermography and breast tumor classification using thermal images." *2020 31st Irish Signals and Systems Conference (ISSC)*. IEEE, 2020.

[21]. Farooq, Muhammad Ali, et al. "Advanced deep learning methodologies for skin cancer classification in prodromal stages." *arXiv preprint arXiv:2003.06356* (2020).

[22]. Khan, Faisal, et al. "Towards Monocular Neural Facial Depth Estimation: Past, Present, and Future." *IEEE Access* (2022).

[23]. Rogalski, Antoni. "History of infrared detectors." *Opto-Electronics Review* 20.3 (2012): 279-308.

[24]. Fitzgerald, Anita, and Jessica Berentson-Shaw. "Thermography as a screening and diagnostic tool: a systematic review." NZ Med J 125.1351 (2012): 80-91.

[25]. Lahiri, Barid Baran, et al. "Medical applications of infrared thermography: a review." *Infrared Physics & Technology* 55.4 (2012): 221-235.

[26]. Wiecek, B. "Review on thermal image processing for passive and active thermography." *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2006.

[27]. Gade, Rikke, and Thomas B. Moeslund. "Thermal cameras and applications: a survey." *Machine vision and applications* 25.1 (2014): 245-262.

[28]. Shirzadfar, Hamidreza, Fatemeh Ghasemi, and Melika Shahbazi. "A Review of Recent Application of Medical Thermography in Human Body for Medical Diagnosis." *SCIOL Biomed* 2 (2018): 102-120.

[29]. Meola, Carosena, Simone Boccardi, and Giovanni Maria Carlomagno, eds. *Infrared thermography in the evaluation of aerospace composite materials: infrared thermography to composites*. Woodhead Publishing, 2016.

[30]. Titman, D. J. "Applications of thermography in non-destructive testing of structures." *NDT & e International* 34.2 (2001): 149-154.

[31]. Jin, S. H. A. O., et al. "Application of infrared thermal imaging technology to condition based maintenance of power equipment [J]." *High voltage apparatus* 49.1 (2013): 126-129.

[32]. Chou, Ying-Chieh, and Leehter Yao. "Automatic diagnostic system of electrical equipment using infrared thermography." 2009 international conference of soft computing and pattern recognition. IEEE, 2009.

[33]. Akula, Aparna, Ripul Ghosh, and H. K. Sardana. "Thermal imaging and its application in defence systems." *AIP conference proceedings*. Vol. 1391. No. 1. American Institute of Physics, 2011.

[34]. Vadivambal, R., and Digvir S. Jayas. "Applications of thermal imaging in agriculture and food industry—a review." *Food and bioprocess technology* 4.2 (2011): 186-199.

[35]. Zhang, Zhuomin M., Benjamin K. Tsai, and Graham Machin. *Radiometric temperature measurements: I. Fundamentals*. Academic press, 2009.

[36]. Diakides, Nicholas A. "New developments in low cost infrared imaging systems." *Eur. J. Thermol* 7.4 (1997): 213-215.

[37]. Balcerak, R., J. P. Jenkins, and N. A. Diakides. "Uncooled focal plane arrays." *Proc. 18th International Conference of IEEE Engineering in Medicine and Biology Society, Amsterdam, Netherlands*. 1996.

[38]. Marshall, Charles A., et al. "Uncooled infrared sensors with digital focal plane array." *Infrared Detectors and Focal Plane Arrays IV*. Vol. 2746. International Society for Optics and Photonics, 1996.

[39]. Rajic, Nik, and Neil Street. "A performance comparison between cooled and uncooled infrared detectors for thermoelastic stress analysis." *Quantitative InfraRed Thermography Journal 11.2 (2014): 207-221*.

[40]. Kirimtat, Ayca, and Ondrej Krejcar. "Flir vs seek in biomedical applications of infrared thermography." *International Conference on Bioinformatics and Biomedical Engineering*. Springer, Cham, 2018.

[41]. FLIR Thermal Homepage, [Online] Available: https://www.flir.eu/" (Last accessed on 21 January 2022).

[42]. SEEK Thermal Homepage, [Online] Available: https://www.thermal.com/compact-series, (Last accessed on 22 January 2022).

[43]. Infratec for medicine, [Online] Available: https://www.infratec.eu/thermography/industries-applications/medicine/, (Last accessed on 22 January 2022).

[44]. Schmal, Martin. *Heterogeneous catalysis and its industrial applications*. Rio de Janeiro: Springer, 2016.

[45]. Pasquini, Celio. "Near infrared spectroscopy: fundamentals, practical aspects and analytical applications." *Journal of the Brazilian chemical society* 14 (2003): 198-219.

[46]. Lynred France, [Online] Available: https://www.lynred.com / (Last accessed on 20th July 2021).

[47]. Budzier, H., and G. Gerlach. "Calibration of uncooled thermal infrared cameras." *Journal of Sensors and Sensor Systems* 4.1 (2015): 187-197.

[48]. Tempelhahn, A., et al. "Improving the shutter-less compensation method for TEC-less microbolometer-based infrared cameras." *Infrared Technology and Applications XLI*. Vol. 9451. International Society for Optics and Photonics, 2015.

[49]. Tempelhahn, Alexander, et al. "Shutter-less calibration of uncooled infrared cameras." *Journal of Sensors and Sensor Systems* 5.1 (2016): 9-16.

[50]. Bieszczad, Grzegorz, and Mariusz Kastek. "Measurement of thermal behavior of detector array surface with the use of microscopic thermal camera." *Metrology and Measurement Systems* 18.4 (2011): 679-690.

[51]. Nvidia TensorRT for developers, [Online] Available, https://developer.nvidia.com/tensorrt, (Last accessed on 14th February 2022).

[52]. Nvidia Jetson Nano, [Online] Available: https://developer.nvidia.com/embedded/jetson-nano-developer-kit, (Last accessed on 14th December 2021).

[53]. Nvidia Jetson Xavier NX Development kit, [Online] Available: https://developer.nvidia.com/embedded/jetson-xavier-nx-devkit, (Last accessed on 14th June 2021).

[54]. Fowler, Kim R. "Automatic gain control for image-intensified camera." *IEEE Transactions on Instrumentation and Measurement* 53.4 (2004): 1057-1064.

[55]. Mudau, Azwitamisi E., et al. "Non-uniformity correction and bad pixel replacement on LWIR and MWIR images." *2011 Saudi International Electronics, Communications and Photonics Conference (SIECPC)*. IEEE, 2011.

[56]. Acconeer radar sensor XM122, [Online] Available https://www.acconeer.com/products, (Last accessed on 14th December 2021).

[57]. Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." *Journal of Big Data* 6.1 (2019): 1-48.

[58]. Goodfellow, Ian, et al. "Generative adversarial networks." *Communications of the ACM* 63.11 (2020): 139-144.

[59]. Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[60]. Feng, Yao, et al. "Joint 3d face reconstruction and dense alignment with position map regression network." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[61]. Raahul, A., et al. "Voice based gender classification using machine learning." *IOP Conference Series: Materials Science and Engineering*. Vol. 263. No. 4. IOP Publishing, 2017.

[62]. Smith, Philip, and Cuixian Chen. "Transfer Learning with Deep CNNs for Gender Recognition and Age Estimation." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.

[63]. Feature Extraction Technique, [Online] Available: https://www.analyticsvidhya.com/blog/2021/04/guide-for-feature-extraction-techniques/, (Last accessed on 08 March 2022).

[64]. CNN application on structured data-Automated Feature Extraction, [Online] Available: https://towardsdatascience.com/cnn-application-on-structured-data-automated-feature-extraction-8f2cd28d9a7e/, (Last accessed on 09 March 2022).

[65]. Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[66]. Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International conference on machine learning*. PMLR, 2019.

[67]. Kelly, Robert. "Critical Comparison of the Classification Ability of Deep Convolutional Neural Network Frameworks with Support Vector Machine Techniques in the Image Classification Process." *Dublin Institute of Technology* (2017).

[68]. Maruyama, Tomoko, et al. "Comparison of medical image classification accuracy among three machine learning methods." *Journal of X-ray Science and Technology* 26.6 (2018): 885-893.

[69]. Yi, Dong, et al. "Learning face representation from scratch." *arXiv preprint arXiv:1411.7923* (2014).

[70]. Kala, Rahul. *On-road intelligent vehicles: Motion planning for intelligent transportation systems*. Butterworth-Heinemann, 2016.

[71]. Munir, Farzeen, et al. "Thermal Object Detection using Domain Adaptation through Style Consistency." arXiv preprint arXiv:2006.00821 2020.

[72]. YOLO-V5 GitHub repository, [Online] Available: https://github.com/ultralytics/yolov5, (Last accessed on 26th February 2021), DOI: 10.5281/zenodo.4154370.

[73]. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[74]. Yolo version camparison, [Online] Available: https://pjreddie.com/darknet/yolo, (Last accessed on 1st February 2022).

[75]. Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934* (2020).

[76]. Davis, James W., and Mark A. Keck. "A two-stage template approach to person detection in thermal imagery." in proceedings of 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1, vol. 1, IEEE, 2005, pp. 364-369.

[77]. Xu, Fengliang, Xia Liu, and Kikuo Fujimura. "Pedestrian detection and tracking with night vision." IEEE Transactions on Intelligent Transportation Systems 6, no. 1, pp: 63-71, 2005, DOI: 10.1109/TITS.2004.838222, [Online].

[78]. Choi, Yukyung, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. "KAIST multi-spectral day/night data set for autonomous and assisted driving." IEEE Transactions on Intelligent Transportation Systems 19, no. 3, 2018, pp: 934-948.

[79]. FLIR Thermal Dataset, [Online] Available: ‘https://www.flir.com/oem/adas/adas-dataset-form/’, (Last accessed on 22nd February 2022).

[80]. Denso Germany, [Online] Available: https://www.denso.com/de/en/, (Last accessed on 15th March 2022).

[81]. Bottou, Léon. "Stochastic gradient descent tricks." *Neural networks: Tricks of the trade*. Springer, Berlin, Heidelberg, 2012. 421-436.

[82]. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014

Appendix A

Proof-of-Concept Techniques for Generating Synthetic Thermal Facial Data for Training of Deep Learning Models

Authors' Contribution to [9]

Contribution Criteria	*Contribution Percentage*
Research Hypothesis	MAF: 70%, PC: 30%
Experiments and Implementation	MAF: 100%
Background	MAF: 100%
Manuscript Preparation	MAF: 80%, PC: 20%

2021 IEEE International Conference on Consumer Electronics (ICCE) | 978-1-7281-9766-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCE50685.2021.9427690

Proof-of-Concept Techniques for Generating Synthetic Thermal Facial Data for Training of Deep Learning Models

Muhammad Ali Farooq
College of Engineering and Informatics
National University of Ireland Galway (NUIG)
Galway, Ireland
m.farooq3@nuigalway.ie

Peter Corcoran
College of Engineering and Informatics
National University of Ireland Galway (NUIG)
Galway, Ireland
peter.corcoran@nuigalway.ie

***Abstract*— Thermal imaging has played a dynamic role in the diversified field of consumer technology applications. To build artificially intelligent thermal imaging systems, large scale thermal datasets are required for successful convergence of complex deep learning models. In this study, we have highlighted various techniques for generating large scale synthetic facial thermal data using both public and locally gathered datasets. It includes data augmentation, synthetic data generation using StyleGAN network, and 2D to 3D image reconstruction using deep learning architectures. Training and validation accuracy of Wide ResNet CNN for binary gender recognition task is improved by 4.6% and 4.4% using original and newly generated synthetic data with an overall test accuracy of 83.33%.**

Keywords—Infrared Imaging, LWIR, Synthetic Data, GAN, Augmentation, Deep Neural Networks, Wide ResNet

I. Introduction

The emerging potential of thermal imaging for consumer applications and the building of low-cost LWIR thermal cameras based on un-cooled sensor technologies is an emerging area of research. As compared to the RGB camera, the thermal camera has its own advantages which includes invariance to illumination changes, an ability to operate even in complete darkness, and provides robust results in the event of shadows and some occlusions. Thermal imaging is used in various Consumer Technology (CT) applications [35] including human thermography for disease diagnosis [1, 2], thermal gender classification [3], human-computer interface systems [4], in-cabin driver and occupant monitoring systems [34].

Modern thermal cameras come with Shutter-less calibration and associated image correction methods such as Non-Uniformity Correction (NUC) and bid pixel replacement are usually applied for real-time processing to produce high-quality thermal sensor data. The resulting thermal images can be employed in a variety of computer vision applications including pedestrian detection, facial recognition [6], object classification [7], and segmentation tasks. However, one drawback of thermal imaging is that there are not many publically available datasets and thus it is challenging to build the large training datasets essentially required to train and fine-tune state-of-art (SoA) Convolution Neural Networks (CNN). In this study, this is addressed by proposing a combination of data augmentation, data generation using Generative Adversarial Networks (GANs), and 2D-3D image reconstruction to enable building substantial numbers of additional synthetic data samples using existing thermal datasets as the seed data.

II. Background

Deep learning models are generally considered as data-hungry models [17]. It requires a vast amount of training data [8] along with proper optimization and regularization techniques to avoid network overfitting and underfitting thus achieving robust results [9]. This problem can be overcome by using data augmentation [12], smart augmentation [14], and synthetically [13] generated datasets. To generate extensive thermal data artificially, from existing datasets, different methods can be used which includes applying various image transformations to the original dataset. These transformations include image rotation, random translations, image cropping, image flipping which additionally includes horizontal and vertical flipping, image shifting, and padding. This method seems to be dynamic and beneficial, not only for the low-data cases but for the imbalance datasets, and also for models trained on large datasets such as Imagenet [15, 16]. Fully synthetic datasets that are generated using computer vision and machine learning tools can be used for diversified computer vision applications such as pose estimation, optical flow, real-time object detection for autonomous driving systems, and text detection [19]. Hu et al. [18] produce novel face images by blending parts of different donor face images. It was done by compositing real facial images thus generating a new set of synthesized data images. In one of the recent studies by Adam, et al [19] authors have used synthetic data for the face recognition task. They have used a 3D morphable facial model for generating images with random amounts of facial identities. In another study, the authors presented Virtual KITTI [20], by using synthetically generated data to train an end-to-end convolutional neural network for object detection, tracking, and scene segmentation tasks for self-driving systems. Similarly, Abbasnejad et al. [21] trained a convolutional neural network using synthetic data for expression analysis tasks. The authors achieved exceptional results in action unit classification on real data. Further, GAN networks are widely used as an image to image translation models such that converting thermal data to synthesized visible data that can be effectively used for training deep learning networks.

III. Implementation Methodology

In this section, we have described the proposed methodology used in this study. Fig. 1 explains the comprehensive block diagram representation for generating large-scale thermal data using existing thermal datasets that can be effectively used for smart thermal imaging systems.

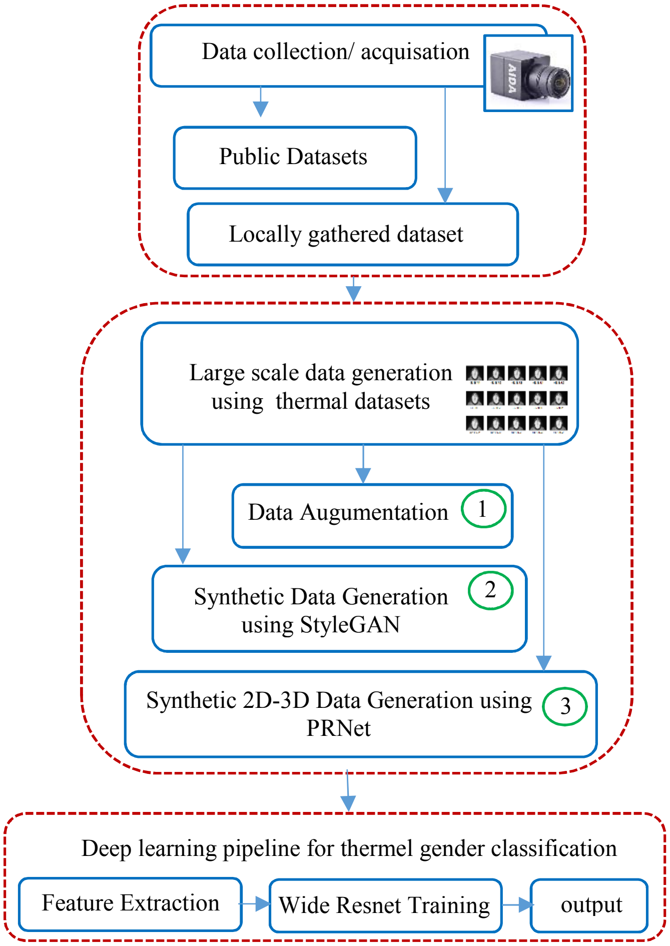

Data collection/ acquisation
Public Datasets
Locally gathered dataset
Large scale data generation using thermal datasets
Data Augumentation
1
Synthetic Data Generation using StyleGAN
2
Synthetic 2D-3D Data Generation using PRNet
3
Deep learning pipeline for thermel gender classification
Feature Extraction
Wide Resnet Training
output

Fig. 1. Block diagram representation of generating large scale thermal data using three different methods.

As shown in Fig. 1 the first step includes data collection and data acquiring using a thermal camera. In this study, we have used three publically available thermal face datasets which include tufts [22-24], carl [27-28], and laval facial thermal dataset [29]. Tufts dataset [22-24] has been published recently with data samples from 6 different image modalities which include visible, near-infrared, thermal, computerized sketch, a recorded video, and 3D images. Moreover, we have gathered our thermal dataset using the LWIR thermal camera which is discussed in the next section of this paper. The recorded thermal data is processed using the non-uniformity correction (NUC) method to adjust the gain and offset for each pixel thus producing a more accurate image. Thermal images have a relatively low signal-to-noise ratio (SNR) [2]. Keeping this in view, sometimes digital image processing techniques are utilized to improve the nature of inferior quality pre-recorded LWIR thermal data. In the second step, we have proposed three different methods for generating thermal synthetic data as shown in Fig. 1. The methods employed in this work are various data augmentation operations, synthetic fake data generation using StyleGAN, and single 2D to 3D thermal image reconstruction using PRNet.

Finally, in the third step, the effectiveness of generated synthetic data has been validated by training state of art Wide ResNet CNN for binary gender classification task using both original and newly generated synthetic data samples.

IV. Applied Methods and Experimental Results for Generating Large Scale Thermal Data

In this section, we have explained various methods used in this study along with their experimental results for generating new thermal data samples using the tufts [22-24], carl [27-28], and laval motion face thermal dataset [29].

A. Data Augmentation

Supervised learning methods require sufficiently large datasets for accurate model training. Image augmentation is considered a beneficial technique in increasing the size of the training set without acquiring new images. It works by bringing supplementary variations in existing data. The generated data samples can be used for robust training of deep convolutional neural networks thus to avoid overfitting, underfitting and better generalize the model for the customized task. There are many different types of image augmentation methods that can be used in accordance with the type of dataset and respective application. Table 1 shows the different types of image augmentation operations available and applied in the Keras framework for performing thermal image augmentation.

TABLE I. Different Types of Image Augmentations

Image Augmentations/ Transformations	
Rotation	Perform image rotation by 30-degree angle.
Flipping	Perform image flipping operations. It includes horizontal flipping and verticle flipping.
Cropping	Perform image cropping at a random location.
Padding	Performs image padding operation with the provided padding value.
Zooming	Performs image zooming with the specified zoom range of 0.15.
Shifting	Image shift is used to add shift-invariance to the images
Affine	Perform affine transformation to the image by provided parameters keeping center invariant.

Fig. 2 shows 9 new training samples generated from a single image of a male and female subject from tufts datasets [22-24] using Table 1 image transformation methods.

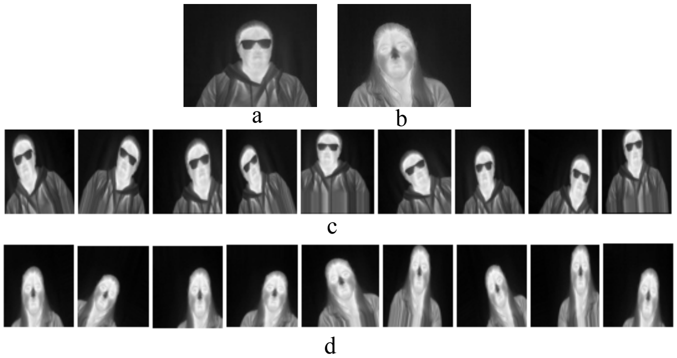


Fig. 2. Image augmentation results: (a) male subject image, (b) female subject image, (c) newly generated male data samples using six different image augmentation methods, and (d) newly generated female data samples using image augmentation.

B. Synthetic Fake Data Genenration using GANs

Generative Adversarial Networks (GANs) are the types of deep neural networks having the capability to generate fake data samples from scratch. The networks work by feeding random noise as the input and once these networks are trained by the selection of proper hyperparameters we can generate realistic data samples. The newly generated fake data samples in the data-limited situation along with original data can be used for optimal training of Convolution Neural Networks (CNN). In this study, we have used StyleGAN [33] for generating synthesized thermal facial samples. StyleGAN [33] is a state of art GAN network introduced by NVIDIA researchers having the ability to generate seemingly infinite numbers of high-resolution data samples. For network training, we have used tufts [22-24], carl [27-28], and laval face motion thermal datasets [29]. Training data comprises varying facial angles, different facial expressions, and subjects with and without glasses. Fig. 3 shows some of the input training samples.

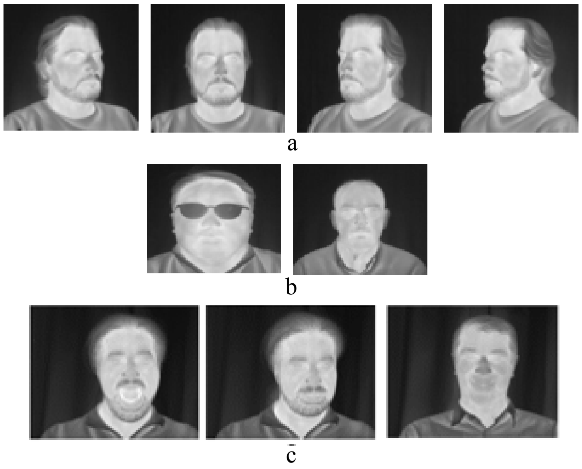


Fig. 3. GAN training data samples (a) varying facial angles of a male subject, (b) four different male subjects with and without glasses, (c) male subjects with different facial expressions (surprise, neutral and happy).

Fig. 4 shows the training structure along with network hyperparameters of StyleGAN architecture.

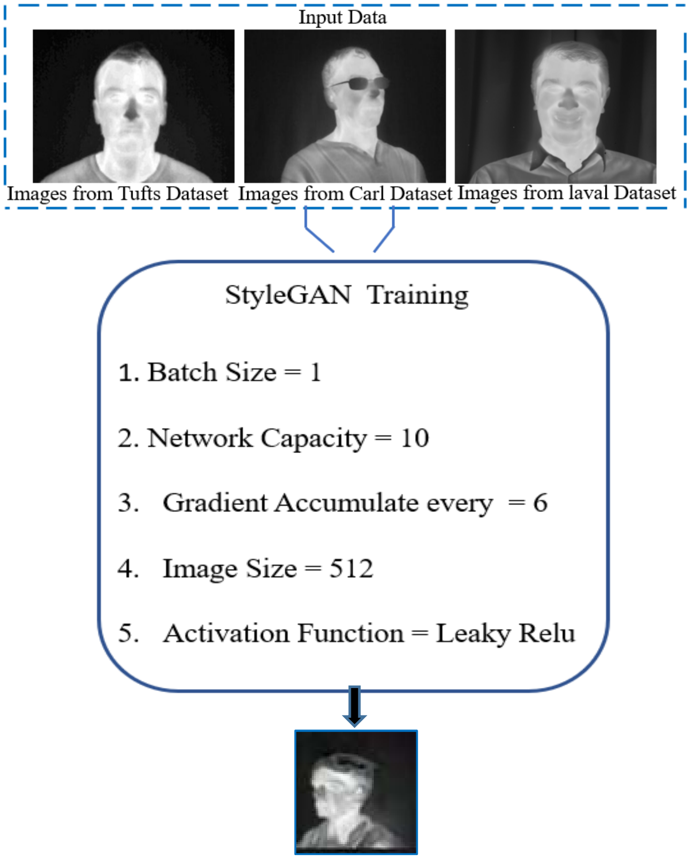

Input Data
Images from Tufts Dataset
Images from Carl Dataset
Images from laval Dataset
StyleGAN Training
1. Batch Size = 1
2. Network Capacity = 10
3. Gradient Accumulate every = 6
4. Image Size = 512
5. Activation Function = Leaky Relu

Fig. 4. StyleGAN training pipeline with a selected set of network hyperparameters.

Fig. 3 and Fig. 4 shows the comprehensive representation of input data and network parameters for generating synthesized thermal outputs. During the training phase, the network was trained for 150,000 epoch with a learning rate of 1e-4. For optimal training of the GAN network, the gradient accumulative mechanism is configured for six steps. It works by splitting the batch of samples into several mini-batches of samples that will run sequentially. The training process is completed in 122 hours with a final generative loss of 0.37 and discriminator loss of 0.24. Fig. 5 shows the intermediate results at different epochs in the form of generated synthetic thermal male facial samples.

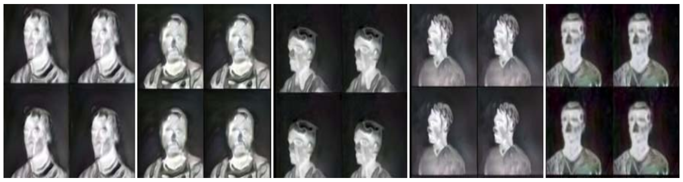


Fig. 5. Synthetic fake thermal facial samples generated using StyleGAN.

Fig. 6 shows some of the generated facial outputs using the trained model in the PyTorch deep learning framework.

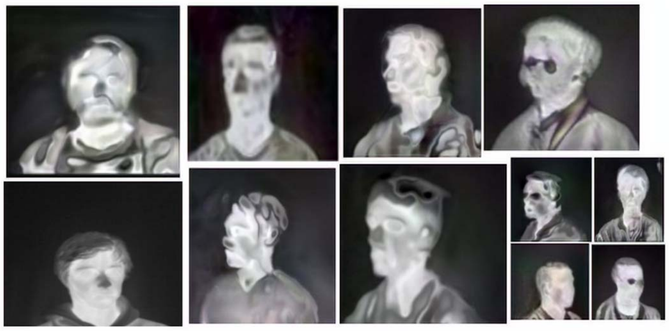


Fig. 6. StyleGAN trained model outputs by generating synthetic thermal face samples.

It can be observed from Fig. 6 that the GAN network generated thermal facial outputs in different variety which includes different facial angles, subjects with and without glasses, different temperature patterns, and different hairstyles. However, results are yet not very robust since facial features are blurred as compared to real thermal facial images. This can be overcome by increasing the training data and last but not least more robust fine-tuning of the GAN networks.

C. 2D to 3D Synthetic Data Genenration using PRNet

This section will present another state of art method from our published paper at Qomex 2020 conference [31]. This method works by generating three-dimensional (3D) synthetic facial geometry structures by employing PRNet [32] on a new set of 2D thermal facial thermal images that are acquired locally. It is an end to end deep learning architecture referred to as Position Map Regression Network (PRN). The system is trained to reconstruct and produce 3D facial images by using a single RGB frontal image of a person [32]. The network works by transferring the input image into a position map. In the next stage, the encoder-decoder structure is used for learning the transfer structure. In this work, we have used the aforementioned network for generating 3D facial geometry structures by using the frontal thermal facial image. The trained network uses Rectified linear Unit (ReLU) activation functions and finally outputs an obj file. The obj file is then imported into blender software for generating 3D thermal facial images covering different facial angles and poses. The same approach is validated on our gathered thermal facial dataset. The data is acquired using an LWIR uncooled 640x480 thermal camera developed under the Heliaus project [5]. The focal length of the camera is 7.5 mm and it has F-number of f/1.2. The prototype thermal camera is shown in Fig. 7 whereas Fig. 8 shows the locally acquired thermal faces of four different subjects in an indoor lab environment.

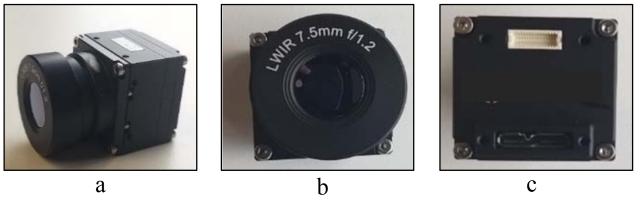


Fig. 7. Prototype thermal VGA camera (a) side view, (b) front view and, (c) back view of the camera.

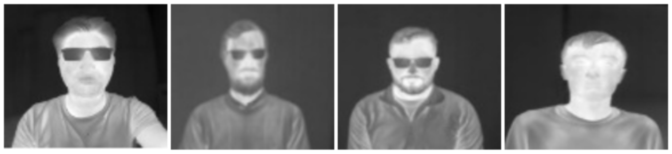


Fig. 8. Thermal image of four different male subjects acquired in an indoor lab environment using uncooled thermal cameras.

Fig. 9 shows the facial depth map along with 3D geometry structures of the male subject generated through PRNet and extracted through blender software using a single frontal frame from Fig. 8.

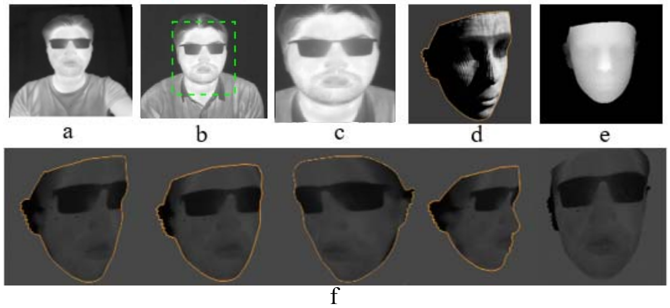


Fig. 9. 3D synthetic facial thermal structures outputs (a) male subject image acquired in an indoor lab environment, (b) Haar cascade face detector, (c) extracted face, (d) 3D facial mesh, (e) 3D depth map generated through PRNet and, (f) different face yaw angles extracted through blender software.

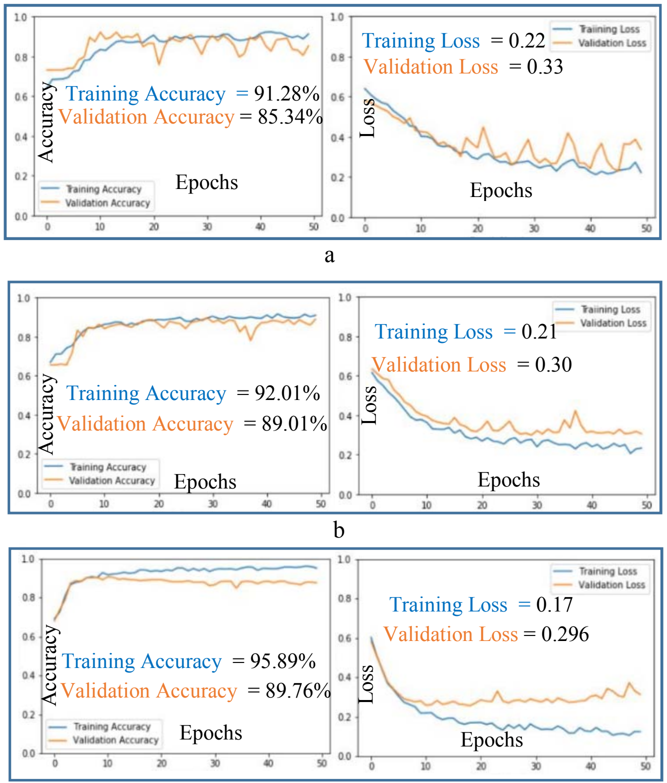
V. Performance Analysis of Deep Learning Model on Original and Synthetically Generated Datasets

In this section, we have analyzed the performance of state-of-the-art deep learning architecture for thermal gender recognition task using transfer learning. For the proposed study we have used Wide ResNet [30] convolution neural network. The wide ResNet architecture is a modified version of originally designed ResNet architecture having an extended number of channels with a total of 68.9 million parameters. Complete experimental techniques for the training of Wide ResNet 50-2 are shown in Table II. The overall training data is divided into a ratio of 80% and 20% for training and validation sets respectively. The model is trained in pytorch deep learning framework on a server-grade machine equipped with 32 GB of Ram and 12 GB TITIAN X graphic card. The model is fine-tuned via transfer learning by adding a few additional layers such that the last FC layer is connected to a linear layer having 256 outputs. It is further fed into the rectified linear unit (ReLU) and dropout layers with the dropout ratio of 0.4 followed by a final FC layer, which has binary output corresponding to the number of classes in tufts dataset.

Fig. 10 shows the training results in the form of accuracy and loss graph of all the Table II experiments. It can be observed from Fig. 10 that experiment 3 got the highest training and validation results as compared to experiment 1 and experiment 2. By taking a close look, we can analyze that the training and validation accuracy of experiment 3 is improved by nearly 4.6% and 4.4% with lower loss values as compared to experiment 1. Thus, we can establish that original thermal images along with processed data, and generated synthetic data, improves the fine-tuning process of the deep learning model. The performance of the best-trained model i.e experiment 3 is cross-validated using the carl thermal facial data along with a locally gathered dataset. The models achieve an overall accuracy of 83.33% on unseen test data.

TABLE II. TRAINING EXPERIMENTS

Wide ResNet 50_2 Training Methods		
No	Training Data Arrangement	Training Parameters
Experiment 1	Model trained on the original tufts dataset.	No of Epochs = 50 Learning Rate = 0.001 Batch size = 32 Drop out = 0.4 Momentum = 0.9 Optimizer = Stochastic Gradient Descent (SGD) Loss Function = Binary Cross-entropy
Experiment 2	Model trained on original tufts dataset along with StyleGAN synthetic data.	
Experiment 3	Model trained using original + StyleGAN synthetic data with image augmentation operations from Table I.	



a

b

c

Fig. 10. Wide ResNet training results (a) Table II experiment 1: training accuracy of 91.28% and validation accuracy of 85.34%, (b) Table II experiment 2: training accuracy of 92.01% and validation accuracy of 89.01%, (c) Table II experiment 3: training accuracy and loss of 95.89 and 0.17 and validation accuracy and loss of 89.76% and 0.29.

VI. CONCLUSION

In this study, we have presented various methodologies for generating synthetic thermal facial samples. The methods employed in this study includes various data transformations, which include rotation, flipping, zooming, cropping, padding, and affine transformation. Secondly, state of art StyleGan architecture is used for generating synthetic fake thermal facial samples. The generated face outputs demonstrate different face angle variations, different hairstyles, image perspective, and with and without glasses. Lastly, we have shown 2D-3D synthetic image reconstruction using a deep learning method by employing state-of-the-art PRNet. The effectiveness of these methods has been evidenced by analyzing improved training results of Wide ResNet Convolution Neural Network (CNN) architecture for the binary gender recognition task. For future work, we can work on other advanced transformation methods such as thermal to visible image translation/ neural style transfer. This can be performed by training advanced GAN architectures which include cycleGAN and pix-to-pix networks thus producing synthetic visible data extracting features from thermal data. Moreover, the efficacy of these methods can be validated on other classification tasks such as thermal facial expression analysis.

ACKNOWLEDGMENT

This synthetic thermal data study for the training of deep learning models to perform gender classification using the public as well locally gathered dataset acquired from prototype thermal camera is part of the project that has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826131. The JU receives support from the European Union's Horizon 2020 research and innovation program and National funding from France, Germany, Ireland (Enterprise Ireland International Research Fund), and Italy.

REFERENCES

[1] Fitzgerald, Anita, and Jessica Berentson-Shaw. "Thermography as a screening and diagnostic tool: a systematic review." NZ Med J 125.1351 (2012): 80-91.

[2] Lahiri BB, S Bagavathiappan, T Jayakumar, John Philip. "Medical applications of infrared thermography: a review". Infrared Physics & Technology 2012;55:221-35.

[3] Chen, Cunjian, and Arun Ross. "Evaluation of gender classification methods on thermal and near-infrared face images." 2011 International Joint Conference on Biometrics (IJCB). IEEE, 2011.

[4] Wettach, Reto, et al. "A thermal information display for mobile applications." Proceedings of the 9th international conference on Human computer interaction with mobile devices and services. 2007.

[5] Heliaus European Union Project Website: https://www.heliaus.eu/, (Last accessed on 20th July2020).

[6] Wu, Zhan, Min Peng, and Tong Chen. "Thermal face recognition using convolutional neural network." 2016 International Conference on Optoelectronics and Image Processing (ICOIP). IEEE, 2016.

[7] Rodin, Christopher Dahlin, et al. "Object classification in thermal images using convolutional neural networks for search and rescue missions with unmanned aerial systems." 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018.

[8] Najafabadi, Maryam M., et al. "Deep learning applications and challenges in big data analytics." Journal of Big Data 2.1 (2015): 1.

[9] Goodfellow, Ian, et al. Deep learning. Vol. 1. Cambridge: MIT press, 2016.

[10] Saha, Arindam, et al. "ThermoFlowScan: Automatic Thermal Flow Analysis of Machines from Infrared Video." *VISIGRAPP (4: VISAPP)*. 2017.

[11] Pretrained deep learning models, Website: https://modelzoo.co/, (Last accessed on 23rd August 2020).

[12] Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." Journal of Big Data 6.1 (2019): 60.

[13] Perez-Ortiz, Maria, et al. "Exploiting synthetically generated data with semi-supervised learning for small and imbalanced datasets." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

[14] Lemley, Joseph, Shabab Bazrafkan, and Peter Corcoran. "Smart augmentation learning an optimal data augmentation strategy." Ieee Access 5 (2017): 5858-5869.

[15] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. IEEE, (2009).

[16] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database.*" 2009 IEEE conference on computer vision and pattern recognition. IEEE*, (2009).

[17] Munappy, Aiswarya, et al. "Data Management Challenges for Deep Learning." 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). IEEE, 2019.

[18] Hu, G., Peng, X., Yang, Y., Hospedales, T.M., Verbeek, J.: Frankenstein: Learning deep face representations using small data. IEEE Transactions on Image Processing 27(1), 293–303 (2018)

[19] Kortylewski, Adam, et al. "Training deep face recognition systems with synthetic data." arXiv preprint arXiv:1802.05891 (2018).

[20] Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. arXiv preprint arXiv:1605.06457 (2016)

[21] Abbasnejad, I., Sridharan, S., Nguyen, D., Denman, S., Fookes, C., Lucey, S.: Using synthetic data to improve facial expression analysis with 3d convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1609–1618 (2017).

[22] Tufts Thermal Dataset, Weblink: http://tdface.ece.tufts.edu/, Last accessed on 29 October 2019.

[23] Panetta, Karen, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Rao et al. "A comprehensive database for benchmarking imaging systems." IEEE Transactions on Pattern Analysis and Machine Intelligence (2018).

[24] Paper: Shreyas Kamath K. M., Rahul Rajendran, Qianwen Wan, Karen Panetta, and Sos S. Agaian "TERNet: A deep learning approach for thermal face emotion recognition", Proc. SPIE 10993, Mobile Multimedia/Image Processing, Security, and Applications 2019, 1099309 (13 May 2019).

[25] Ebner, Marc. Color constancy. Vol. 7. John Wiley & Sons, 2007.

[26] Ancuti, Cosmin, et al. "Enhancing underwater images and videos by fusion." *2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE*, 2012.

[27] V. Espinosa-Duró, M. Faundez-Zanuy and J. Mekyska, "A New Face Database Simultaneously Acquired in Visible, Near-Infrared and Thermal Spectrums*", Cognitive Computation, vol. 5, no. 1, pp. 119-135*, 2013.

[28] V. Espinosa-Duró, M. Faundez-Zanuy, J. Mekyska and E. Monte-Moreno, "A Criterion for Analysis of Different Sensor Combinations with an Application to Face Biometrics", *Cognitive Computation, vol. 2, no. 3, pp. 135-141*, 2010

[29] Ghiass, Reza Shoja. "Face Recognition Using Infrared Vision." Ph.D. diss., Université Laval, 2014.

[30] Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks." arXiv preprint arXiv:1605.07146 (2016).

[31] M. A. Farooq and P. Corcoran, "Generating Thermal Image Data Samples using 3D Facial Modelling Techniques and Deep Learning Methodologies," , Twelfth International Conference on Quality of Multimedia Experience (QoMEX), Athlone, Ireland, pp. 1-5, 2020.

[32] Feng, Yao, et al. "Joint 3d face reconstruction and dense alignment with position map regression network." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[33] Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[34] Ebrahimian-Hadikiashari, S., et al. "Monitoring the variation in driver respiration rate from wakefulness to drowsiness: a non-intrusive method for drowsiness detection using thermal imaging." J Sleep Sci 3.1-2 (2018): 1-9.

[35] Vollmer, Michael, and Klaus-Peter Möllmann. Infrared thermal imaging: fundamentals, research and applications. John Wiley & Sons, 2017.

Appendix B

Generating Thermal Image Data Samples using 3D Facial Modelling Techniques and Deep Learning Methodologies

Authors' Contribution to [10]

Contribution Criteria	*Contribution Percentage*
Research Hypothesis	MAF: 70%, PC: 30%
Experiments and Implementation	MAF: 100%
Background	MAF: 100%
Manuscript Preparation	MAF: 70%, PC: 30%

2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)

Generating Thermal Image Data Samples using 3D Facial Modelling Techniques and Deep Learning Methodologies

Muhammad Ali Farooq
College of Engineering and Informatics
National University of Ireland Galway (NUIG)
Galway, Ireland
m.farooq3@nuigalway.ie

Peter Corcoran
College of Engineering and Informatics
National University of Ireland Galway (NUIG)
Galway, Ireland
peter.corcoran@nuigalway.ie

***Abstract*—Methods for generating synthetic data have become of increasing importance to build large datasets required for Convolution Neural Networks (CNN) based deep learning techniques for a wide range of computer vision applications. In this work, we extend existing methodologies to show how 2D thermal facial data can be mapped to provide 3D facial models. For the proposed research work we have used tufts datasets for generating 3D varying face poses by using a single frontal face pose. The system works by refining the existing image quality by performing fusion based image preprocessing operations. The refined outputs have better contrast adjustments, decreased noise level and higher exposedness of the dark regions. It makes the facial landmarks and temperature patterns on the human face more discernible and visible when compared to original raw data. Different image quality metrics are used to compare the refined version of images with original images. In the next phase of the proposed study, the refined version of images is used to create 3D facial geometry structures by using Convolution Neural Networks (CNN). The generated outputs are then imported in blender software to finally extract the 3D thermal facial outputs of both males and females. The same technique is also used on our thermal face data acquired using prototype thermal camera (developed under Heliaus EU project) in an indoor lab environment which is then used for generating synthetic 3D face data along with varying yaw face angles and lastly facial depth map is generated.**

***Keywords*— thermal, CNN, synthetic, deep learning, 2D, 3D, LWIR**

I. Introduction

With the recent advancements in technology and the growing requirements for larger datasets, it is very important to extract maximum information from the acquired data. Visible or RGB data is most commonly used for a wide range of computer vision applications however it is not able to generate temperature patterns of the specific body which is an important factor in critical applications. Thermal cameras can capture the temperature patterns of the human body by sensing the emission of infrared radiation. Thermal data of the human body is considered to be very important for many applications such as human disease diagnosis in early stages by extracting human facial and body temperature patterns and medical image analysis techniques and creating 3D synthetic thermal face data for visualization and animations. In the proposed study we have proposed 3D synthetic thermal face data generation by using advanced deep learning methods inspired by Feng, Yao, et al [1]. Such types of data can be used in different types of human biometric applications such as thermal facial recognition systems, gender classification system, emotion recognition systems.

As compared to 2D facial images 3D facial structures can help in dealing with the problem of varying human poses and occlusions. Deep learning algorithms have played a vital role in solving many computer vision applications including 3D data creation by taking advantage of convoluting neural networks (CNN). CNN is well known for its self-feature extraction from the raw pixel of images rather than relying on handcrafted features which are subsequently required for conventional machine learning algorithms.

The rest of the paper is structured as follows, section II provides the background and related research for creating synthetic data, section III describes the proposed methodology regarding image refinement and deep learning for generating the 3D facial structures, section IV provides experimental results and lastly, section V describes the conclusion and future work.

II. Background

Synthetic data can be considered as a repository of data that is not collected from real-world experiments, but it is generated programmatically by using different algorithms and methodologies from the domain of machine learning and pattern recognition. The most common approach for generating synthetic data is to pick the work of 3D artists they have done by creating real-time virtual environments for video gaming. In [2] authors have captured datasets from the Grand Theft Auto V video game. The authors mainly emphasize on using semantic segmentation methods. Authors have captured the communication between the game and the graphics hardware. Through this approach, they have cut the labeling costs (in annotation time). Weichao Qiu and Alan L. Yuille [4] have developed UnrealCV which is an open-source plugin for the popular game engine Unreal Engine 4. It works by providing commands that allow us to get and set camera location, field of view and get the set of objects in a scene together with their positions. C. Choi and H. I. Christensen in [6] created a dataset of 3D models of household objects for their tracking filter. Moreover, Hodan et al. [7] provide a real dataset of textureless objects supplemented with 3D models of these objects. Deep learning has been extensively used for generating 3D data especially biometrics data for various real-world applications. Dou, Pengfei, Shishir K. Shah, and

978-1-7281-5965-2/20/$31.00 ©2020 IEEE

Ioannis A. Kakadiaris [8] and Tuan Tran, Anh, et al [9] have proposed an end to end CNN architectures to directly estimate the 3D morphable models (3DMM) shape parameters. In this study, we have proposed a novel method to generate a comprehensive 3D thermal facial structure by using state of art CNN architecture.

III. Proposed Methodology

This section of the paper will mainly focus on the proposed algorithm for generating 3D synthetic face data from a single 2d thermal image. We have utilized the tufts thermal face dataset [10-12] since this dataset has published recently with data samples from 6 different image modalities which include visible, near-infrared, thermal, computerized sketch, a recorded video, and 3D images. The dataset consists of images of both males and females genders. It was acquired in an indoor environment using FLIR Vue Pro Camera with constant lighting. Fig. 1 displays sample thermal images of both male and female subjects from the tufts dataset [10-12].

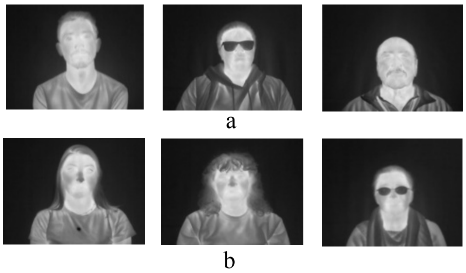


Fig. 1. Thermal face images from tufts dataset a) male samples, b) female samples

In the first phase, the system works by taking a single frontal pose and producing the refined version of the image to make facial features such as eyes, lips, nose and temperature patterns on the face more visible and vibrant. This approach works by applying various image preprocessing operations built on multi-scale fusion principles inspired by Ancuti, Cosmin, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert Ancuti [13].

A. *2D Image Processing*

The algorithm consists of six main steps. In the first step, algorithms work by applying a simple white color balance which is color corrections operation to remove the unlikely color casts in order to render specific colors in an image. The resulting outputs are color corrected version with reduced noise levels as compared to original raw data. In the second stage, Contrast limited Adapt Histogram Equalization (CLAHE) is applied to enhance the visibility of confined details by improving the contrast of local regions in the image. It is done to achieve optimal contrast levels in the input image. In the next stage, different types of weighs are applied to increase the exposure levels in the dark regions. This is achieved by applying four different types of weights which include laplacian contrast weights, local contrast weights, saliency contrast weights, and exposedness weights [13]. Laplacian weights are generally used to enhance the global contrast of the image. Local contrast weights are applied to strengthen the local contrast appearance since it advantages the transitions mainly in the highlighted and shadowed parts of images. It is computed as the standard deviation between pixel luminance level and the local average of its surrounding region as shown in equation 1[13].

$$W_{lc}\,(x,y) = II(I^k - I^k_{Whc})II \qquad (1)$$

Where $W_{lc}\,(x,y)$ represent the symbol for local contrast weights, I^k represents the luminance channel of the input and the I^k_{Whc} represents the low-passed version of it.

Saliency weights are applied to emphasize the discriminating objects that lose their prominence especially in the dark regions and exposedness weights are finally applied to evaluate how well the pixel is exposed. The weights are measured by using the saliency algorithm of Achanta et al.[18]. Fig. 2 represents the complete multi-scale fusion image refinement process on the thermal frontal face image from the tufts dataset [10-12].

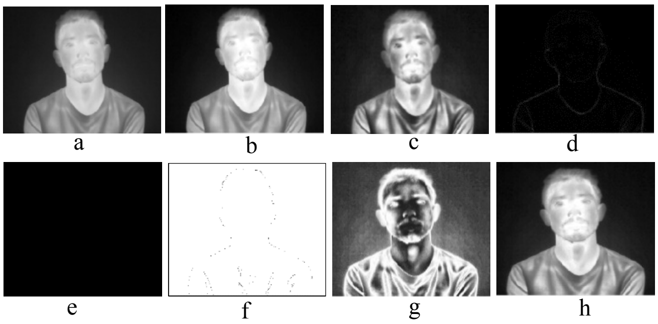


Fig. 2. Complete multi-scale image fusion algorithm pipeline to produce a refined version of an image a) input image, b) white color balance applied, c) histogram equalization applied, d) laplacian contrast weight applied, e) local contrast weights applied, f) saliency weights applied, g) exposedness weights applied, h) final output image

B. *2D to 3D Image Reconstruction*

Once the images are refined in the second stage, we have used end to end convolution neural network also referred to as Position Map Regression Network (PRN) [1] to reconstruct the 3D images from a single frontal face pose thermal image. The authors in [1] had proposed the CNN network which was trained to generate 3D facial structures using one single RGB image.

The network works by transferring the input image into a position map. In the next stage, the authors have used the encoder-decoder structure for learning the transfer structure. The encoder structure is consisting of one convolution layer which is followed by a series of ten residual blocks for performing downsampling operation. The decoder structure is consisting of seventeen transposed convolutions blocks in order to generate the predicted output position map. The proposed CNN networks use Rectified linear Unit (ReLU) activation functions and a kernel size of four is used for each of the convolution layers and transposed convolution layers. A customized loss function was built to learn the parameters to a better extent by measuring the difference between the ground truth position map and the network output. The loss function utilizes the weight mask which is the grey image recording the weight of each map on the position map. The weigh mask is of the same size and pixel to pixel correspondence when compared with the position map. The loss function is defined in equation 2 [1].

This project (https://www.heliaus.eu/) has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826131. The JU receives support from the European Union’s Horizon 2020 research and innovation program and France, Germany, Ireland, Italy.

$$Loss\ = \sum \text{II}\ Pos\ (u,v) - Pos{\sim}(u,v)\ \text{II}\,.\,\mathrm{W}(\mathrm{u},\mathrm{v}) \qquad (2)$$

Where $Pos\ (u,v)$ represent predicted position map, $Pos{\sim}(u,v)$ represents ground truth position map and $\mathrm{W}(\mathrm{u},\mathrm{v})$ represents the weight mask.

In this study, we have used the same network for generating synthetic 3D facial geometry structures using one single image. However, instead of using the visible image we have utilized thermal facial images to validate the effectiveness of the network. The network initially produces .obj file which is then imported in blender software [14] to generate the 3D facial geometry as the output. The complete workflow diagram is shown in Fig. 3

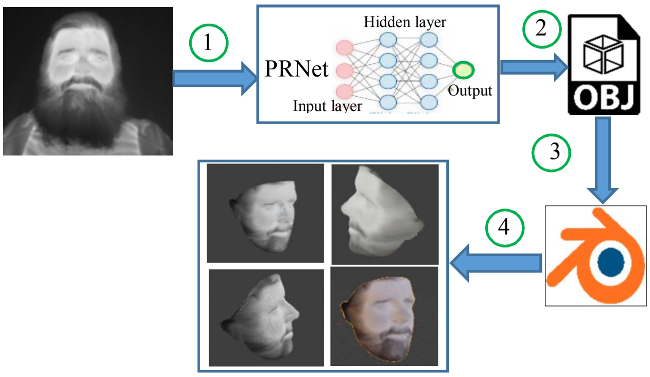


Fig. 3. Complete workflow diagram for generating the synthetic 3D facial structure from single 2D thermal image 1: input image fed to PRNet for generating 3D facial data, 2: output obj file, 3: obj file imported to blender software, 4: final outputs extracted in the form of 3D thermal facial images covering different poses

IV. Experimental Results

The overall algorithm was implemented using Matlab R2018a for applying fusion based image preprocessing methods as discussed in Section III to produce the refined version of images. TensorFlow [15] deep learning framework was used for generating 3D facial structures using the pre-trained PRNet [1]. The system was deployed and tested on the Core I7 machine with 32 GB of RAM equipped with NVIDIA RTX 2080 Graphical Processing Unit (GPU) having 8GB of dedicated graphic memory.

The first phase of the experimental results shows the refined outputs obtained by applying fusion based image preprocessing operations. It makes the facial features and temperature patterns of the face more visible by reducing the overall noise and adjusting the optimal brightness and contrast levels in the image. It is shown in Fig. 4.

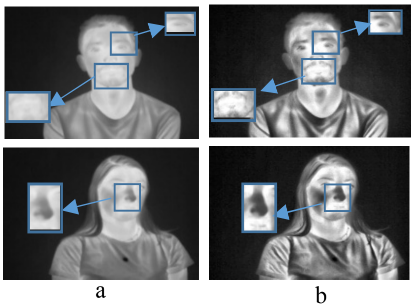
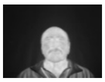
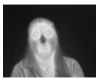


Fig. 4. Image preprocessing results a) input images of two different subject (male and female), b) refined outputs with more visible facial features

In the proposed study we have used different image quality metrics which include Naturalness Image Quality Evaluator (NIQE) [16] and Blind/Reference less Image Spatial Quality Evaluator (BRISQE) [17] score to compare the enhanced version of images with original (ground-truth) images. It is shown in Table I.

TABLE I. Image Quality Metrics

Image	NIQE Score	BRISQUE Score
Processed Image	**3.0323** lower is better	**18.2226** lower is better
Orignal Image	3.3350	37.2575
Processed Image	**2.9099** lower is better	**29.4339** lower is better
Orignal Image 	3.6029	35.7555
Processed Image	**2.7059** lower is better	**12.0921** lower is better
Orignal Image 	3.4460	33.5439

The main reason for employing these two metrics is that it does not require any reference image. Therefore, these metrics are also known as no-reference or objective-blind image quality analyzers. The lower NIQE and BRISQUE scores of processed images (bolded) reflect that image quality is improved significantly by applying the image refinement/ preprocessing techniques. The second phase of the experimental results demonstrates the 3D facial structures generated through PRNet [1] and extracted through blender software using a refined version of the thermal image. It is shown in Fig. 5

The same process is used to produce varying 3D facial poses of both male and female samples from the tufts dataset [10-12]. It is shown in Fig. 6.

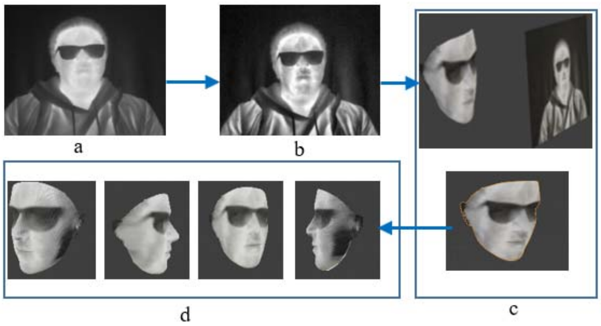


Fig. 5. Image preprocessing results a) input images of two different subject (male and female), b) refined outputs with more visible facial features c) 3D face geometry generated in blender software, d) different face poses of the subject

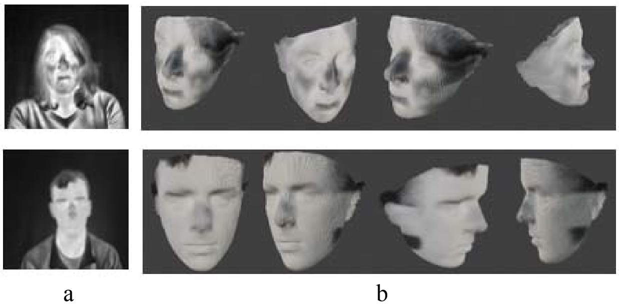


Fig. 6. 3D synthetic face structure of both male and female gender a) refined input images of female and male, b) synthetic 3D outputs with varying face poses

Further, the second phase of experimental results demonstrate 3D facial structures using our dataset. The data is gathered in the indoor lab environment using a prototype thermal camera that embeds a Lynred [20] LWIR sensor developed under the Heliaus EU project [19]. Fig. 7 displays the prototype thermal camera model being used for the proposed research work whereas Table II provides the technical specifications of the camera.

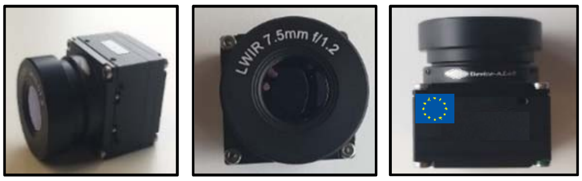


Fig. 7. Prototype thermal VGA camera model for acquiring local data

TABLE II. TECHNICAL SPECIFICATIONS OF PROTOTYPE THERMAL CAMERA

Specifications	
Type	Long Wave Infrared (LWIR)
Resolution	640 x 480 pixels
Quality	VGA
Focal length	7.5 mm

The data is collected by mounting the camera on a tripod stand with the fixed distance (nearly 60 centimeters) from the subject The data acquisition structure is shown in Fig. 8.

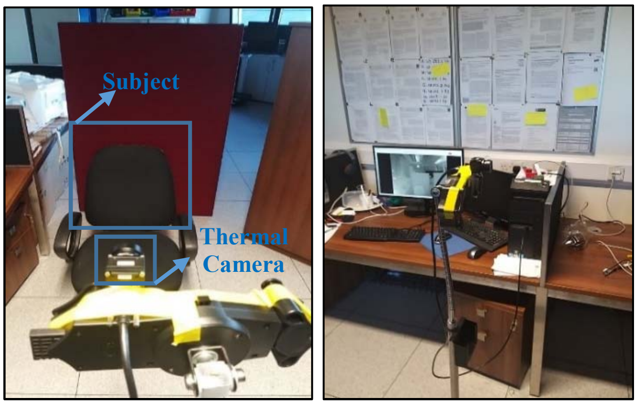


Fig. 8. Data acquisition setup in the indoor university lab environment, prototype thermal camera mounted on tripod stand placed at a fixed distance (nearly 60 centimeters) from the subject

We have collected data in two different modalities which include RGB and thermal respectively. For preliminary testing, only male subjects took part in this study. The data is gathered by recording the video and then extracting the image sequence from the video. Fig. 9 displays frontal face poses of two different subjects referred to as subject A and subject B who have taken part in this study along with their processed thermal outputs.

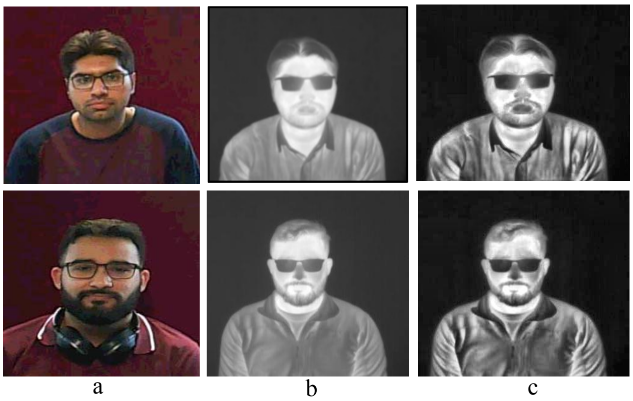


Fig. 9. Face data samples of two different subjects (first-row subject A, second-row subject B) acquired in an indoor lab environment a) visible image, b) thermal image, c) processed thermal image

After collecting the data, the same technique is used to generate 3D facial structures. Along with 3D facial structures, we have also generated a facial depth map. It is exhibited in Fig. 10.

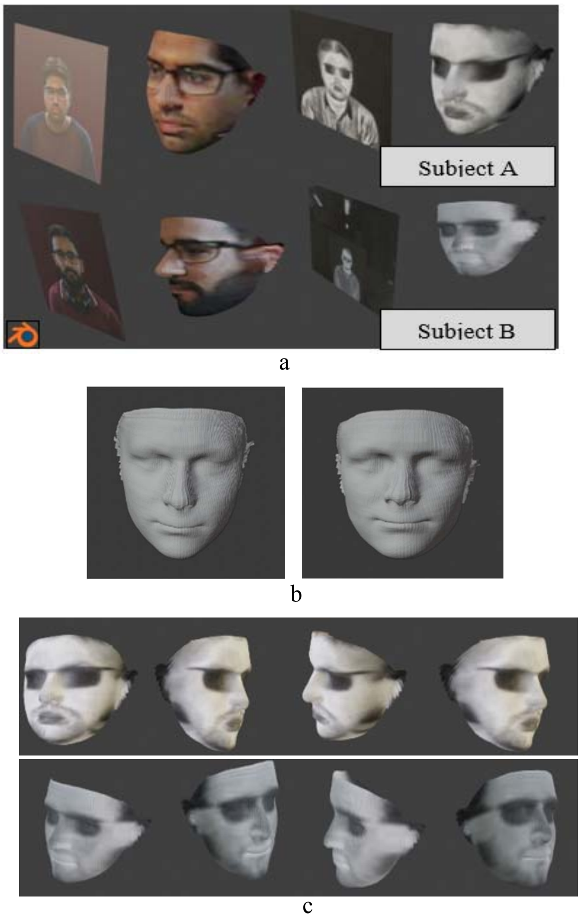


Fig. 10. Synthetic 3D facial structure results with depth map on our own data a) RGB and thermal Obj file imported in Blender software, b) 3D facial mesh of subject A and subject B, c) different face yaw angles of subject A and subject B

Lastly, we have generated the facial depth map on our collected lab data using the PRNet [1]. The main goal of generating the facial depth map is to provide more comprehensive and reliable 3D information which can be further used for facial recognition systems in conjunction with different 3D facial poses. It is shown in Fig. 11.

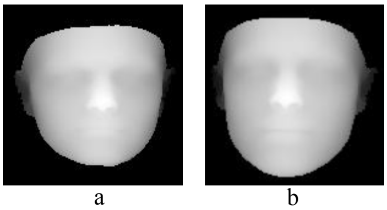


Fig. 11. Face depth map generated a) subject A facial depth map, b) subject B facial depth map

V. Conclusion and Future Work

In the proposed study we have incorporated advance deep learning model PRNet [1] for generating synthetic 3D thermal facial structures from the single thermal frontal image. Such type of data can be found useful in a variety of real-time computer vision and machine learning applications such as medical image analysis, extracting facial and body temperatures for in-cabin driver monitoring systems, visualization, and animation creation. As compared to conventional data generation techniques such as data augmentation and data transformation that can produce 2D outputs, 3D synthetic data can be found more robust and realistic especially when training deep neural networks for critical applications. Along with thermal and visible data, we can use this methodology to create synthetic data from other image modalities such as infrared, near-infrared and grayscale data. Moreover, as future work, we can train CNN networks to generate 3D facial textures that can be aligned with 3D mesh for generating animations, different facial expressions and facial emotions with varying lighting conditions, and occlusions.

Acknowledgment

This research is supported and funded by the Heliaus European Union Project. The project focused on enabling safe autonomous driving systems. This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826131. The JU receives support from the European Union's Horizon 2020 research and innovation program and France, Germany, Ireland, Italy. The authors would like to acknowledge Shubhajit Basak for providing his support to use blender software, the Xperi Ireland team and Quentin Noir from Lynred France for giving their feedback. Moreover, authors would like to acknowledge tufts university the contributors of the tufts dataset for providing the image resources to carry out this research work.

References

[1] Feng, Yao, et al. "Joint 3d face reconstruction and dense alignment with position map regression network." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[2] Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. CoRR, abs/1709.07322, 2017.

[3] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. CoRR, abs/1608.02192, 2016.

[4] Weichao Qiu and Alan L. Yuille. Unrealcv: Connecting computer vision to unreal engine. CoRR, abs/1609.01326, 2016.

[5] Weichao Qiu, Fangwei Zhong, Yi Zhang, Zihao Xiao Siyuan Qiao, Tae Soo Kim, Yizhou Wang, and Alan Yuille. Unrealcv: Virtual worlds for computer vision. ACM Multimedia Open Source Software Competition, 2017.

[6] C. Choi and H. I. Christensen. Rgb-d object tracking: A particle filter approach on gpu. In 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 1084–1091, Nov 2013.

[7] Tomas Hodan, Pavel Haluza, Step´an Obdrz´alek, Jiri Matas, Manolis I. A. Lourakis, and Xenophon Zabulis. T-LESS: an RGB-D dataset for 6d pose estimation of texture-less objects. CoRR, abs/1701.05498, 2017.

[8] Dou, Pengfei, Shishir K. Shah, and Ioannis A. Kakadiaris. "End-to-end 3D face reconstruction with deep neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[9] Tuan Tran, Anh, et al. "Regressing robust and discriminative 3D morphable models with a very deep neural network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[10] Tufts Thermal Dataset, Weblink: http://tdface.ece.tufts.edu/, Last accessed on 29 October 2019.

[11] Panetta, Karen, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Rao et al. "A comprehensive database for benchmarking imaging systems." IEEE Transactions on Pattern Analysis and Machine Intelligence (2018).

[12] Paper: Shreyas Kamath K. M., Rahul Rajendran, Qianwen Wan, Karen Panetta, and Sos S. Agaian "TERNet: A deep learning approach for thermal face emotion recognition", Proc. SPIE 10993, Mobile Multimedia/Image Processing, Security, and Applications 2019, 1099309 (13 May 2019).

[13] Ancuti, Cosmin, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert. "Enhancing underwater images and videos by fusion." In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 81-88. IEEE, 2012

[14] Blender software Website: https://www.blender.org/, (Last accessed on 10th February 2020).

[15] TensorFlow Deep Learning Platform, Website: https:// www.tensorflow.org, (Last accessed on 8th February 2020).

[16] Mittal, Anish, Anush Krishna Moorthy, and Alan Conrad Bovik. "No-reference image quality assessment in the spatial domain." IEEE Transactions on image processing 21.12 (2012): 4695-4708.

[17] Mittal, Anish, Rajiv Soundararajan, and Alan C. Bovik. "Making a "completely blind" image quality analyzer." IEEE Signal Processing Letters 20.3 (2012): 209-212.

[18] Achantay, R. "Hemamiz., S.; Estraday, F. Frequency-tuned salient region detection." Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR), Miami, FL, USA. 2009.

[19] Heliaus European Union Project Website: https://www.heliaus.eu/, (Last accessed on 20th January 2020).

[20] Lynred France Website: https://www.lynred.com/, (Last accessed on 27th January 2020).

Appendix C

Performance estimation of the state-of-the-art convolution neural networks for thermal images-based gender classification system

Authors' Contribution to [11]

Contribution Criteria	*Contribution Percentage*
Research Hypothesis	MAF: 80%, HJ: 20%
Experiments and Implementation	MAF: 100%
Background	MAF: 100%
Manuscript Preparation	MAF: 70%, HJ: 15%, PC: 15%

Performance estimation of the state-of-the-art convolution neural networks for thermal images-based gender classification system

Muhammad Ali Farooq,[a,*] **Hossein Javidnia**,[a,b] **and Peter Corcoran**[a]

[a]National University of Ireland Galway, College of Engineering and Informatics,
Galway, Ireland
[b]ADAPT Centre, Trinity College, Dublin, Ireland

Abstract. Gender classification has found many useful applications in the broader domain of computer vision systems including in-cabin driver monitoring systems, human–computer interaction, video surveillance systems, crowd monitoring, data collection systems for the retail sector, and psychological analysis. In previous studies, researchers have established a gender classification system using visible spectrum images of the human face. However, there are many factors affecting the performance of these systems including illumination conditions, shadow, occlusions, and time of day. Our study is focused on evaluating the use of thermal imaging to overcome these challenges by providing a reliable means of gender classification. As thermal images lack some of the facial definition of other imaging modalities, a range of state-of-the-art deep neural networks are trained to perform the classification task. For our study, the Tufts University thermal facial image dataset was used for training. This features thermal facial images from more than 100 subjects gathered in multiple poses and multiple modalities and provided a good gender balance to support the classification task. These facial samples of both male and female subjects are used to fine-tune a number of selected state-of-the-art convolution neural networks (CNN) using transfer learning. The robustness of these networks is evaluated through cross validation on the Carl thermal dataset along with an additional set of test samples acquired in a controlled lab environment using prototype uncooled thermal cameras. Finally, a new CNN architecture, optimized for the gender classification task, GENNet, is designed and evaluated with the pretrained networks. [DOI: 10.1117/1.JEI.29.6.063004]

Keywords: deep convolution neural networks; thermal imaging; gender classification; long-wave infrared; transfer learning.

Paper 200318 received May 1, 2020; accepted for publication Oct. 23, 2020; published online Nov. 18, 2020.

1 Introduction

Uncooled thermal imaging is approaching a level of maturity where it can be considered as an alternative to, or as a complimentary sensing modality to that of visible or NIR imaging. Thermal imaging offers some advantages as it does not require external illumination and provides a very different perspective on an imaged scene than a conventional CMOS-based image sensor. The proposed research work is carried under HELIAUS[1] project, which is focused on in-cabin driver monitoring systems using thermal imaging modality. The driver gender classification in a vehicle can help to improve the personalization of various features (e.g., user interfaces and presentation of data to the driver). It can also be used to better predict driver cognitive response,[2] driver behavior, and intent, and finally knowledge of gender can be useful for safety systems such as airbag deployment that may adapt to driver physiology. In summary, automotive manufacturers are interested to have the knowledge of driver gender within the vehicular environment for designing smarter and safer vehicles. Alongside this, there are many other applications of thermal human gender classification systems. In security systems, thermal imaging can easily detect people and animals even in total darkness. In human–computer interaction systems, thermal

[*] Address all correspondence to Muhammad Ali Farooq, m.farooq3@nuigalway.ie

imaging can provide complimentary information, determining subtle fluctuations in facial temperatures that can inform on the emotional status of a subject. In other human–computer interaction systems, the systems may need to classify the individual person and/or their facial expressions and voices[3] in order to effectively interact with them thus gender information serves as a source of soft biometrics.[4] In medical applications, human thermography provides an imaging method to display heat emitted from a human body surface thus helping us to understand unique facial thermal patterns in both male and female gender.[5] Human thermography helps us to better understand that central and peripheral thermoreceptors are distributed all over the body including on the human face and are responsible for both sensory and thermoregulatory responses to maintain thermal equilibrium. Studies have shown that heat emission from the surface of the body is symmetrical. All these studies measured differences between the left and right side of different areas of the head.[6,7,8]

The literature reports that in healthy subjects the difference in skin temperature from side to side of the human body is as small as 0.2°C.[8] The heat emission from the human body is related to cutaneous vascular activity, yielding enhanced heat output on vasodilation, and reduced heat amount on vasoconstriction.[9] The medical literature reports that a significant difference has been observed between the absolute facial skin temperature of men and women during the clinical studies of facial skin temperature.[9] Men were found to have higher temperatures compared to women overall; 25 anatomic areas were measured on the face including upper lips, lower lips, chin, orbit, and the cheek. According to another study, the basal metabolic rate of a healthy 30-year-old male with a height of 5 ft, 7 in weight of 64 kg, and who has surface area of about 1.6 m^2 dissipates about 50 $\mathrm{W/m^2}$ of heat; on the other hand the basal metabolic rate of healthy 30-year-old female with the height of 5 ft, 3 in the weight of 54 kg, and who has surface area of 1.4 $\mathrm{W/m^2}$ dissipates about 41 $\mathrm{W/m^2}$ of heat. In addition, women's skin is expected to be cooler since less heat is lost per unit of body surface area.[9] However, thermal patterns whether in the case of male or female also depend on many other factors such as age, human body intrinsic and extrinsic characteristics, outdoor environmental conditions, and technical factors such as camera calibration, and the field of view (FoV). Moreover, it also depends on factors such as drinking, smoking, various diseases, and using medications.

The preliminary focus of this study is on binary human gender classification, however, the same system can be retrained for third or multi-class (non-binary) gender classification tasks if such datasets are available.

In this study, the Tufts thermal faces[10–12] and Carl thermal faces datasets[13,6] are used to train and test a selection of state-of-the-art neural networks to perform the gender classification task. Figure 1 shows some examples of thermal facial images with varying poses from the Tufts dataset and frontal facial poses from the Carl dataset. The complete workflow pipeline is detailed in Sec. 3 of this paper. In addition to using pretrained neural networks, a new CNN architecture, GENNet, is provided. This is designed and trained specifically for the gender classification task and is evaluated against the pretrained CNN networks. In addition, a new validation set of thermal images is acquired in controlled laboratory conditions using a new prototype uncooled thermal camera and is used as a second means of cross-validating all the pretrained models along with GENNet architecture. The evaluation results are presented in Sec. 4.

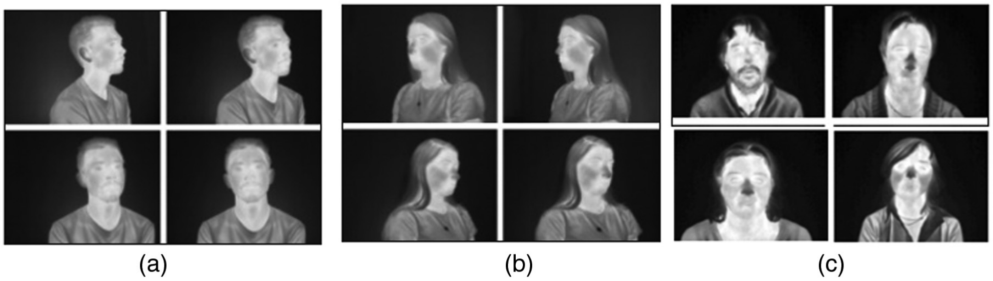


Fig. 1 Sample images from Tufts and Carl thermal face database: (a) male subject with four different face poses from the Tufts dataset; (b) female subject with four different face poses from the Tufts dataset; and (c) male and female subjects (frontal face pose) from Carl database.

2 Background/Related Work

This section focuses on the background research and previous studies on gender classification using CNNs.

2.1 *Gender Classification Using Conventional Machine Learning Methods*

Makinen and Raisamo[14] and Reid et al.[15] provided a detailed survey of the gender classifications method in their studies. One of the early techniques for gender recognition reported in Ref. 16 utilized a neural system trained on a small arrangement of close frontal face pictures. In Ref. 17, the consolidated 3D structure of the head (captured by a laser scanner) and picture intensities were utilized for characterizing genders. Support vector machine (SVM) classifiers were employed by Ref. 18 where the authors evaluated the performance of SVM with an overall error rate of 3.4% when compared with other traditional classifiers including linear, fisher linear discriminant, nearest neighbor, and radial basis functions. Instead of using SVM,[19] Baluja and Rowley[20] referred to AdaBoost for gender classification tasks using a set of low-resolution grayscale images. Perspective invariant age and gender recognition was performed by Ref. 21 using arbitrary viewpoints. Recently, Ullah et al.[22] utilized the Webers local surface descriptor[23] for the gender recognition system, showing near-perfect execution on the facial recognition technology (FERET) benchmark.[24] In Ref. 25, shape, texture, and color features were extracted from frontal faces, thus obtaining robust outcomes on the FERET benchmark. In an attempt by Arun and Rarath,[26] unique mark pictures are used, and the input images are represented by a feature vector consisting of ridge thickness to valley thickness ratio and ridge density. Further, they used SVM to categorize subjects into male and female classes accordingly. In addition to the gender classification system using the visible spectrum, the possibility of deducing gender information from thermal and NIR spectrum is also gaining much interest. Chen and Ross[27] claimed to be the first proposing human faces-based gender classification system using thermal and NIR data. The authors have selected three different conventional feature extraction methods for gender representation including linear binary patterns, principle component analysis, and pixels from low-resolution facial images. For gender recognition, they have used SVM, LDA, Adaboost, random forest, Gaussian mixture model, and multi-layer perceptron classifiers. Their experimental results conclude that SVM for histogram-based gender classification results in much better performance on NIR and thermal spectra. Nguyen and Park[28] proposed a gender classification system using joint visible and thermal spectrum data of the human body. The classification accuracies in Ref. 28 are measured by employing different feature extractors including HoG and MLBP.[29] Their experimental results demonstrated an improvement in classification accuracy using the joint data from visible and thermal image spectrums. Similarly, in another study reported in Ref. 30, the author's utilized multimodal datasets consisting of audiovisual, thermal, and physiological recordings of male and female subjects. The authors extracted feature values from these datasets, which were later used for automatic gender classification purposes. In both studies, authors used conventional machine learning algorithms for feature extraction rather than using advanced deep learning methodologies.

2.2 *Gender Classification Using Deep Learning-Based Methods*

Due to the fact that much potential is laid in deep CNN structures, they are widely used for diversified applications especially where more precise and robust accuracy levels are required such as medical image analysis, surveillance systems, object detection, and autonomous classification systems.[31] Canziani et al.[32] listed many pretrained models that can be used for various practical applications in their study. They analyzed the overall performance of these pretrained models by computing the accuracy levels and the inference time needed for each model. Dwivedi and Singh[33] provided a comprehensive review of deep learning methodologies for robust gender classification using the GENDER-FERET[34] face dataset. In their study, they have compared the performance of various CNN architectures. Moreover, they have selected one of the architectures as a baseline model, and by changing different parameters like the number of fully connected (FC) layers and the number of filters they have created different models. The authors achieved the best accuracy of 90.33% with the base model architecture of CNN. Ozbulak et al.[35]

have investigated two different deep learning strategies including fine-tuning and SVM classification using CNN features. They were applied on different networks including their proposed task-specific GilNet model and pretrained domain-specific VGG[36] and Generic AlexNet[37]-like CNN model for building robust age and gender classification system using the Adience[38] visible spectrum dataset. The experimental results from their study show that transferred models outperform the GilNet model for both age and gender classification tasks by 7% and 4.5%, respectively. In a more recent study, Manyala et al.[39] investigated the overall performance of two CNN-based methods for gender classification using near-infrared (NIR) images. In the first method, a pretrained VGG-Face[40] was used for extracting features for gender classification from a convolutional layer in the network, whereas the second method used a CNN model obtained by fine-tuning VGG-Face to perform gender classification from periocular images. The authors had achieved the classification accuracy of 81% on an in-house dataset, which was gathered locally.

Further in a more recent study, Baek et al.[41] used the combined data of both visible and NIR spectrum for performing robust gender classification using full human body images in surveillance environment. The system works by deploying two CNN architecture to remove the noise of visible-light images and enhance the existing image quality to improve gender recognition accuracy. The overall system performance was evaluated on desktop pc as well as on Jetson TX2 embedded system.

3 Research Methodology

The goal of this work is to evaluate the potential of thermal image facial data as a means of gender classification. The thermal image data are analyzed with a selected set of nine state-of-the-art neural networks. These pre-existing convolution neural networks are adapted for the thermal data using transfer learning. In addition, a new CNN model is proposed, and its performance is compared against nine state-of-art pretrained networks.

Initially, all the pretrained networks are first trained on the Casia Face dataset[42] since Tufts thermal training dataset[10–12] does not contain enough images, an important requirement for optimal training of deep neural networks. This face dataset is used to extract low-level features for building the baseline architecture. In the second stage, the Tufts thermal face database[10–12] is used for transfer learning. This dataset consists of 113 different subjects and comprises images from six different image modalities that include visible, NIR, thermal, computerized sketch, a recorded video, and 3D images of both male and female classes. The thermal face dataset was acquired in a controlled indoor environment using constant lighting that was maintained using diffused lights. Thermal images were captured using FLIR Vue Pro Camera,[43] which was mounted at a fixed distance and height.

Figure 2 represents the complete workflow diagram of the overall gender classification system.

3.1 *Initial Training and Transfer Learning of Pretrained Networks*

This research takes advantage of the pretrained networks by freezing and unfreezing all the layers and adding customized final layers to generalize the model for the target autonomous gender classification task from thermal image datasets. The main reason for using these pretrained networks is they already learned low-level feature values such as edges and textures by training the networks on very large and varied datasets. This process helps in obtaining useful results even with a relatively small training dataset since the basic image features have already been learned by the pretrained model using larger datasets like ImageNet.[44] Further, the classifier is trained to learn the higher-level features in the proposed thermal dataset images.

A typical CNN system comprises certain layers including convolution layers, pooling layers, dense layers, and FC layers. There are various pretrained networks available that can be efficiently used for different types of visual recognition, object detection, and segmentation tasks. For the proposed study, the following pretrained neural networks are utilized: ResNet-50,[45] ResNet-101,[45] Inception-V3,[46] MobileNet-V2,[47] VGG-19,[36] AlexNet,[37] DenseNet-121,[48] DenseNet-20,[48] and EfficientNet-B4[49] networks. These models are chosen as they are commonly

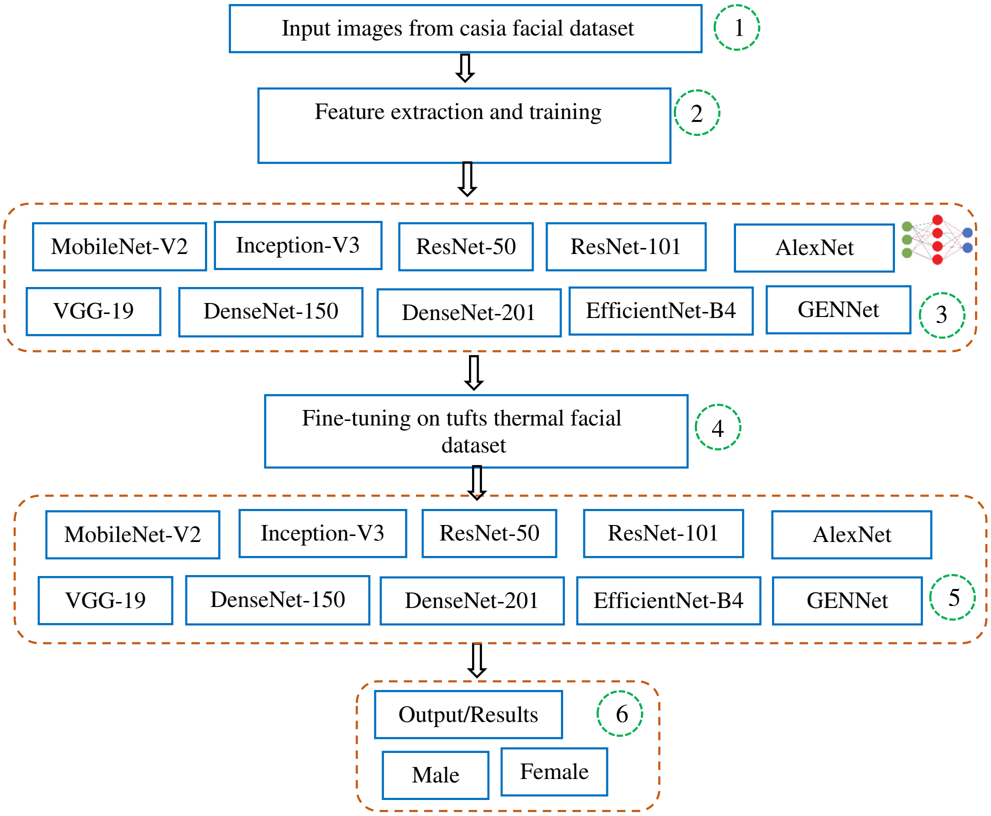


Fig. 2 Workflow diagram for autonomous gender classification system using thermal images.

trained using the ImageNet[44] dataset, each model has a different architectural style, they provide a good trade-off between accuracy and inference time,[50] and in addition, they are the state-of-the-art for image classification tasks. Thus an impartial performance comparison of these networks can be made for the thermal gender classification task.

ResNet[45] architecture mainly relies on the residual learning process. The network is designed to solve complex visual tasks using more deeper layers stacked together. ResNet-50 is a 50-layer Residual Network. The other variants from the ResNet family include ResNet-101[45] and ResNet-152.[45] Resnet-50 network was initially trained on ImageNet,[44] which consists of a total of 1.28 million images from 1000 different classes. The Inception-v3 is made up of 48 layers stacked on top of each other.[46] The Inception-v3 model was initially trained on Imagenet[44] as well. These pretrained layers have a strong generalization power as they are able to find and summarize information that will help to classify various classes from the real-world environment.

MobileNet-V2 is considered as efficient deep learning architecture proposed by Sandler et al.[47] specifically designed for mobile and embedded vision applications. It is a lightweight deep learning architecture with the working principle of using depth-wise separable convolutions meaning that it performs a single-convolution operation on each color channel rather than combining all three and flattening them. This has the advantage of filtering the input channels.

DenseNet[48] architecture also referred to as dense convolutional neural network is a state-of-the-art variable-depth deep convolutional neural architecture. It was designed to improve the architecture of ResNet.[45] The principle design feature of this architecture is channel-wise concatenation, with every convolution layer that has access to the activations of every layer preceding it. DenseNet family has different variants including DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264.

VGGNet[36] was developed by the Visual Geometry Group from the University of Oxford. Like ResNet[45] and Inception-V3,[46] this network was also originally trained on ImageNet.[44] The network was designed with the significant improvement compared to AlexNet architecture,[37] which was more focused on smaller window sizes and strides in the first convolutional

Table 1 Performance comparison of state-of-the-art CNN

CNN	Number of parameters	Top 5 error rate	Depth	Main attributes
AlexNet	62 M	ImageNet: 16.4	8	Uses ReLU, dropout, and overlap pooling
VGGNet	138 M	ImageNet: 7.3	19	Homogenous topology, uses small size kernels
Inception-V3	24 M	ImageNet: 3.5	159	Replace large size filters with small filters
MobileNet	2.2 M	ImageNet: 10.5	17	The width multiplier uniformly reduces the number of channels at each layer, fast inference
ResNet-50	26 M	ImageNet: 3.6	152	Residual learning, identity mapping-based skip connection
ResNet-101	43 M			
DenseNet-121	7.2 M	CIFAR-10+: 3.46	190	Cross-layer information flow
DenseNet-201	18.6 M			
EfficientNet-B4	19 M	ImageNet: 2.9		Compound coefficient scaling method, 8.4 × smaller and 6.1 × faster than other convnets

layer. VGG architecture can be trained using images with (224×224) pixel resolution. The main attribute of VGG architecture is that it uses very small receptive fields (3×3 with a stride of 1) compared to AlexNet[37] (11×11 with a stride of 4). In addition to this, VGG incorporates 1×1 convolutional layers to make the decision function more non-linear without changing the receptive fields. The architectures come in different variants including VGG-11, VGG-16, and VGG-19.

EfficientNet[49] was recently published and designed using a compound scaling method. As the name suggests the network proved to be a competent and optimum network by achieving state-of-the-art results on the ImageNet dataset. Table 1[51] provides a more comprehensive comparison of these architectures highlighting their attributes, number of parameters, the overall error rate on benchmark datasets, and their respective depth.

As discussed in the previous section, all the pretrained networks are initially trained on the Casia Face database[42] since the Tufts thermal training dataset[10–12] does not contain a sufficient number of images. Casia facial dataset[42] consists of facial images of different celebrities (38,423 distinct subjects) in the visible spectrum. This facial dataset has been used to extract low-level feature values for building a baseline architecture. The networks are trained using a total of 30,887 frontal facial images of different celebrities from both genders. The data were split in the ratio of 90% for training and 10% for validation. To better generalize and regularize the base model for final fine-tuning on the thermal dataset, certain data transformations are performed on the Casia[42] training data including random resizing of 0.8, random rotation of 15 deg, and flipping. The logic for performing these transformations is that it will bring supplementary data variations for optimal training of the baseline architectures keeping in view the final fine-tuning process on thermal images. Figure 3 displays the Casia data samples along with training data transformation results. The initial training is done by adding a small number of additional final layers to enable generalization and regularization of all the pretrained models. In the case of ResNet-50 and ResNet-101 networks, the last FC layer is connected to a linear layer having 256 outputs. It is further fed into the rectified linear unit (ReLU)[52] and dropout layers with the dropout ratio of 0.4 followed by a final FC layer, which has binary output corresponding to the two classes in the Casia dataset. A similar formation of final layers is inserted by transforming the number of features to the number of classes in all the pretrained networks. Each of these networks is further fine-tuned using a training dataset comprising of thermal facial image samples. The fine-tuning is achieved using transfer learning techniques.[53]

The models were trained using the PyTorch framework.[54] Binary cross-entropy is used as the loss function during training along with a stochastic gradient descent (SGD)[55] optimizer. The final training data include male and female thermal images as shown in Fig. 4.

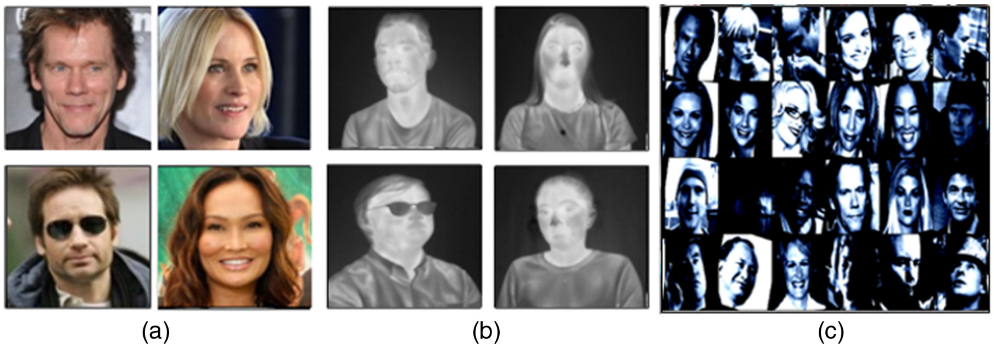


Fig. 3 Facial samples from two different datasets: (a) male and female data samples from Casia[42] database; (b) male and female samples from Tufts thermal images;[10–12] and (c) PyTorch data transformations on Casia dataset.

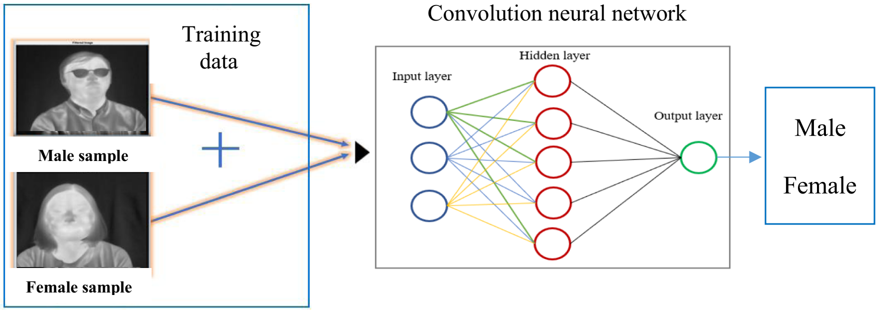


Fig. 4 Training data comprising of male and female samples for network training.

In order to better fine-tune the networks, the thermal training data are augmented by introducing a selection of image variations. These are achieved using the transformation operations shown in Table 2.

During the fine-tuning phase, the SGD[55] and the Adam[56] optimizers are used to compare their respective performance. This is discussed in Sec. 4. As compared to gradient descent (GD) where the full training set is used to update the weights in each iteration, in minibatch SGD,[55] the dataset is split into randomly samples minibatches, and the weights are updated in separate iterations for each minibatch (not element-wise unless minibatch size is 1). Moreover, minibatch SGD[55] is computationally less expensive and minimizes losses faster than GD as it cycles through the full training data, just in the form of chunks as opposed to all at once. The Adam[56] optimizer is an adaptive learning rate optimizer and is considered one of the best optimizers for training convolution neural networks. As compared to minibatch SGD, Adam optimizer also uses the SGD algorithm. However, it implements an adaptive learning rate and

Table 2 Training data transformation

Transformation type	Data variation

Resized cropping	Size = 256,
	scale = (0.8, 1.0)

Rotation	15 deg
Flipping	Horizontal
Center cropping	Size: 224
Tensor conversion	—
Mean and standard deviation normalization	[0.485, 0.456, 0.406], [0.229, 0.224, 0.225]

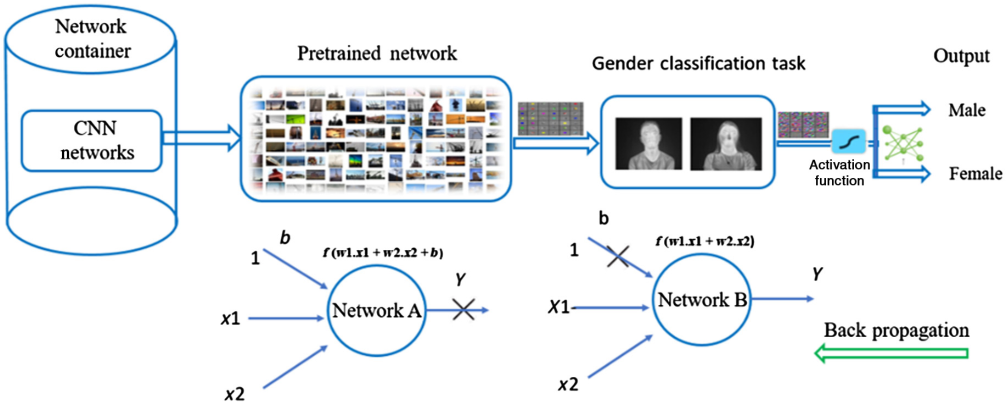


Fig. 5 CNN training structure: network A indicates pretrained networks with initial weights and network B indicates transfer learning process with new weights for thermal gender classification.

Table 3 Pretrained networks hyperparameters

Network hyperparameters	
Batch size	32
Epochs	100
Learning rate	0.001
Momentum	0.9
Loss function	Cross-entropy
Optimizer	SGD and Adam

can determine an individual learning rate for each parameter. Figure 5 shows the generalized training structure for all the pretrained networks. The training data are split into the ratio of 80% and 20% for training and validation purposes, respectively. To achieve a fair evaluation baseline, all the pretrained networks are fine-tuned using the same hyper-parameters on the one train dataset. These parameters are provided in Table 3.

3.2 *New CNN Model GENNet*

To analyze the validity of the existing thermal images, a novel CNN network is designed that is referred to as GENNet and its performance is compared against the pretrained state-of-the-art architectures. The structural block diagram representation of the proposed network is shown in Fig. 6. The overall network structure is consisting of four main blocks. The first three blocks

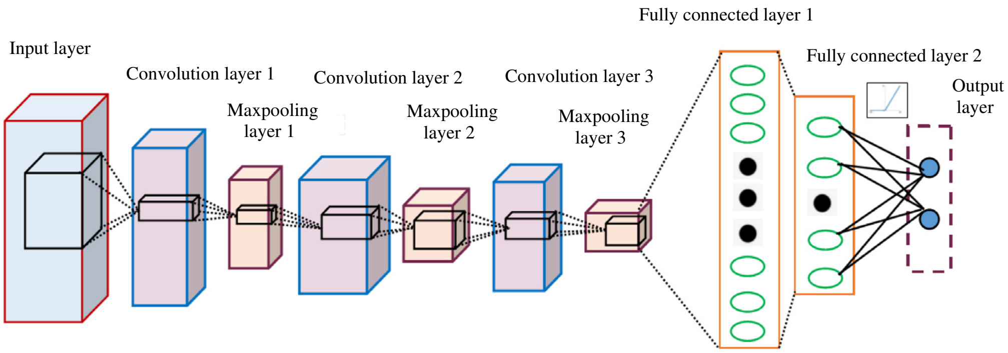


Fig. 6 Structural representation of GENNet CNN model for thermal images-based gender classification.

contain sequential layers in the form of 2D convolutions each followed by the ReLU[52] activation function, max-pooling, and dropout layers. The fourth block consists of two FC layers. The first FC layer is followed by the ReLU activation function[52] and dropout layer, whereas the second and last FC layer of the overall network converts the corresponding number of features to the number of outputs. The layer-wise detail of the GENNet model is provided in Appendix A (Table 7).

Like all other pretrained networks, GENNet is initially trained on the Casia facial database[42] and later fine-tuned on Tufts thermal dataset.[10–12] The same division of thermal training data is used along with the same hyperparameters as it was utilized for other pretrained models. Once the network is fine-tuned, it is tested on the combination of two new datasets as discussed in Sec. 4.3.

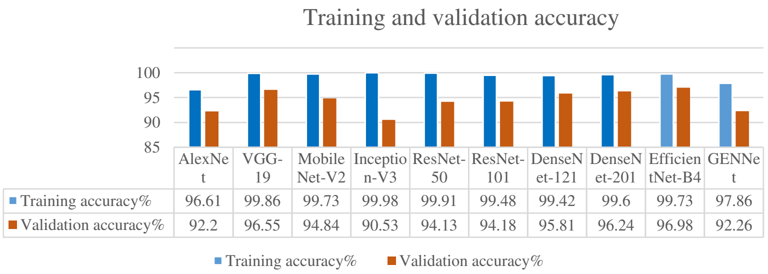
4 Experimental Results

PyTorch[54] deep learning platform is used to fine-tune and train all the pretrained models as well as the proposed GENNet model. These experiments are performed on a machine equipped with NVIDIA TITAN X graphical processing unit with 12 GB of dedicated graphic memory.

4.1 *Training and Validation Results of CNN Architectures by Unfreezing the Layers*

In this part of the experimental study, all the networks are retrained by unfreezing all the original network layers to improve the feature learning process on thermal data. As described and shown in ablation study Sec. 6, transfer learning while freezing the network layers and using both SGD and ADAM optimizer we cannot achieve optimal training and validation accuracy in the case of most of the models. The experimental results using freezed network layer are depicted in Fig. 14. During this fine-tuning process, both Adam and SGD optimizers were employed and the best results in the case of each model were selected. Most of the models performed well, achieving better training and validation accuracy as shown in Fig. 7. AlexNET is specifically trained using a fixed learning rate and it utilizes a one-cycle learning policy to achieve a better convergence. The initial learning rate of the network is set to 0.001 and momentum to 0.9. The final learning rate of the network was 0.0003. Using a smaller learning rate makes a model converge more efficiently but at the expense of the speed, whereas using a higher learning rate can lead to model divergence. Thus to overcome this issue, the learning rate needs to be adjusted automatically. One cycle LR works by increasing and then decreasing the learning rate according to a fixed schedule during the complete training process of a CNN. The main goal of performing these techniques is to optimize all the models as well as that of the newly proposed GENNET architecture. Figure 7 shows the training and validation accuracy chart of all the retrained networks along with the newly proposed GENNet architecture.

It can be observed that most of the models performed significantly well by getting training accuracy above 96% and validation accuracy greater than 90%. The inception-V3 achieved the highest training accuracy with the lowest training loss of 0.008. The Efficientnet-B4 network achieved the highest validation accuracy of 96.98% with a validation loss of 0.11. The newly proposed GENNet model for task-related thermal gender classification achieves the overall training and validation accuracy of 97.86% and 92.26% with loss of 0.08 and 0.15, respectively. The trained models are further used for cross-validating their performance on the new test data as discussed and shown in the subsections.



	AlexNet	VGG-19	MobileNet-V2	Inception-V3	ResNet-50	ResNet-101	DenseNet-121	DenseNet-201	EfficientNet-B4	GENNet
■ Training accuracy%	96.61	99.86	99.73	99.98	99.91	99.48	99.42	99.6	99.73	97.86
■ Validation accuracy%	92.2	96.55	94.84	90.53	94.13	94.18	95.81	96.24	96.98	92.26

Fig. 7 Accuracy charts of all the networks by unfreezing the network layers.

4.2 *Local Thermal Data Acquisition*

To further validate the effectiveness of all the pretrained models and provide an additional mode of comparison with the newly proposed CNN GENNet model, a live thermal facial dataset was gathered using a new prototype thermal camera. The data are acquired in an indoor lab environment using a camera-based on a prototype uncooled microbolometer thermal camera array that embeds a Lynred[57] long-wave infrared (LWIR) sensor developed under the Heliaus EU project.[1] Figure 8 displays the prototype thermal camera model being used for the proposed research work to gather this live dataset, whereas Table 4 provides the technical specifications of the camera.

To take comprehensive facial information during the data acquisition process, we have calculated other important parameters including the lens aperture, angular field of view (AFOV), height and width of the sensor, and working distance as shown as follows:[58]

$$F-\text{number} = \frac{\text{focal length}(f)}{\text{diameter}(D)}, \tag{1}$$

$$\text{diameter}(D) = \frac{\text{focal length}(f)}{F\,\text{number}} = \frac{7.5}{1.2} = 6.25 \approx 6\ \text{mm}, \tag{2}$$

$$\text{height of sensor}(h) = \text{horizontal pixels} * \text{pixel spitch} = 640 * 17 = 10.88\ \text{mm}, \tag{3}$$

$$\text{width of sensor}(w) = \text{vetricle pixels} * \text{pixel spitch} = 480 * 17\ \mu\text{m} = 8.16\ \text{mm}, \tag{4}$$

$$\text{AFOV} = 2 * \tan^{-1}\frac{h}{2f} = 2 * \tan^{-1}\frac{10.88\ \text{mm}}{2 * 7.5\ \text{mm}} = 71.9 \approx 72\ \ \text{deg}, \tag{5}$$

$$\text{working distance(WD)} = \frac{\text{focal length}(f) * \text{HFOV}}{\text{height of sensor}(h)} = \frac{7.5 * 890}{10.88} \approx 60\ \text{cm}. \tag{6}$$

The data are collected by mounting a camera on a tripod at a fixed distance of 60 to 65 cm. The height of the camera is adjusted manually to align the subject's face centrally in the FoV. Shutterless[59] camera calibration at 30 FPS is used to acquire the data. The data acquisition setup

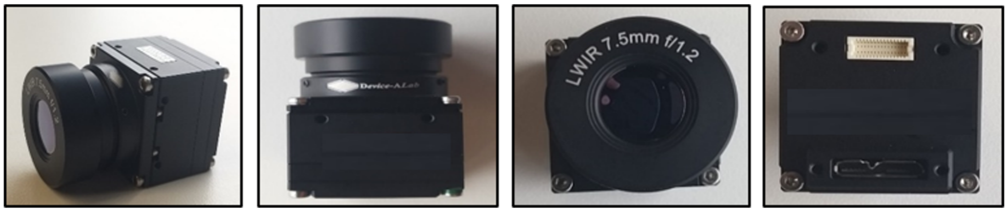


Fig. 8 Prototype thermal VGA camera model for acquiring local facial data.

Table 4 Technical specifications

Prototype thermal camera specifications	
Quality and type	VGA and LWIR
Resolution	640 × 480 pixels
Focal length (*f*)	7.5 mm
F-number	1.2
Pixel pitch	17 μm
HFOV	90 deg, 890 mm

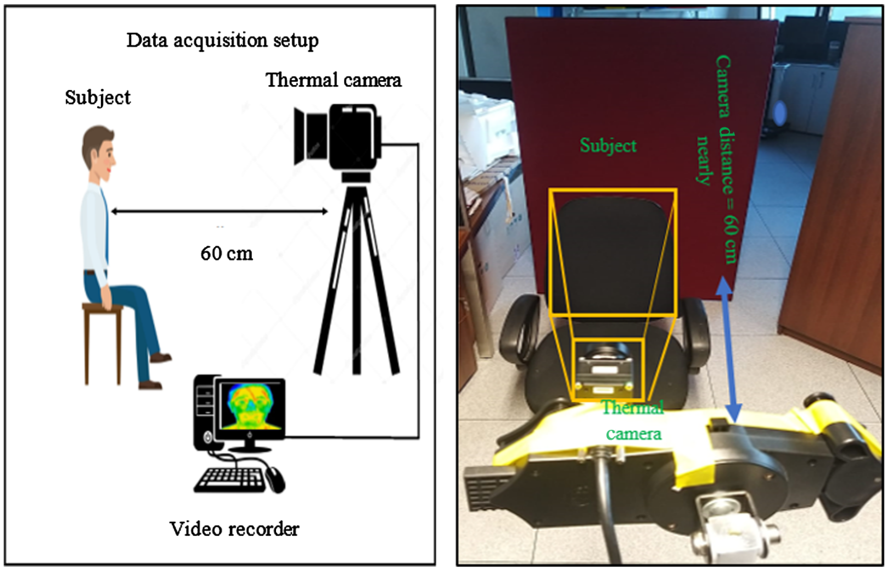


Fig. 9 Indoor lab environment data acquisition setup.

is shown in Fig. 9. A total of five subjects consensually agreed to take part in this study. The data were gathered by recording videos stream of each subject covering different facial poses and then generating image sequences from the acquired videos.

Figure 10 illustrates a few samples of the captured data including both male and female subjects.

4.3 *Testing Results of State-of-the-Art CNN*

All the trained models are tested on the combination of the two different datasets including Carl[13,6] and the locally gathered indoor thermal dataset. This is done to cross-validate the effectiveness of all the trained classifiers, as discussed in Sec. 1. The best models achieving the highest training and validation accuracy from Sec. 4.3 are selected for the cross-validation experiment. The test data contain a total of ninety samples. The overall performance of all the networks on test data is measured using the accuracy metric as shown in the following equation:[60]

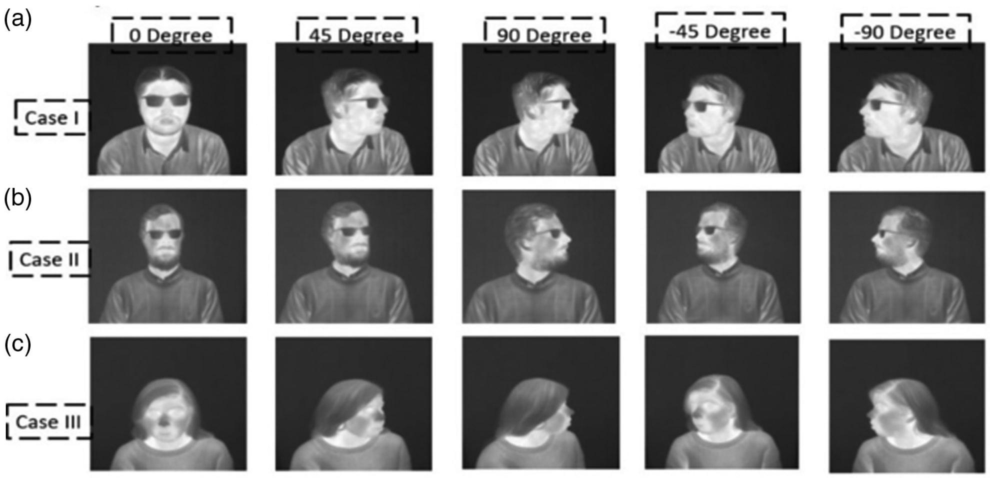
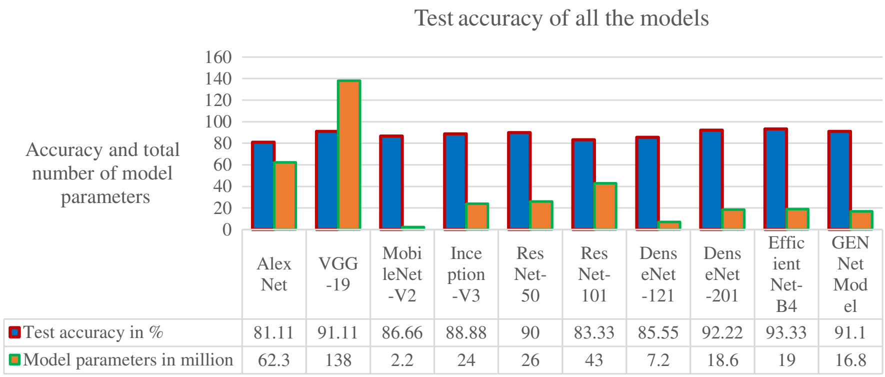


Fig. 10 Test cases of three different subjects acquired in the lab environment with varying face pose: (a), (b) the varying facial angles of male subjects and (c) the different facial angles of a female subject.



	Alex Net	VGG -19	Mobi leNet -V2	Ince ption -V3	Res Net- 50	Res Net- 101	Dens eNet -121	Dens eNet -201	Effic ient Net- B4	GEN Net Mod el
🟦 Test accuracy in %	81.11	91.11	86.66	88.88	90	83.33	85.55	92.22	93.33	91.1
🟧 Model parameters in million	62.3	138	2.2	24	26	43	7.2	18.6	19	16.8

Fig. 11 Test accuracy and model parameters chart of all the CNN architectures.

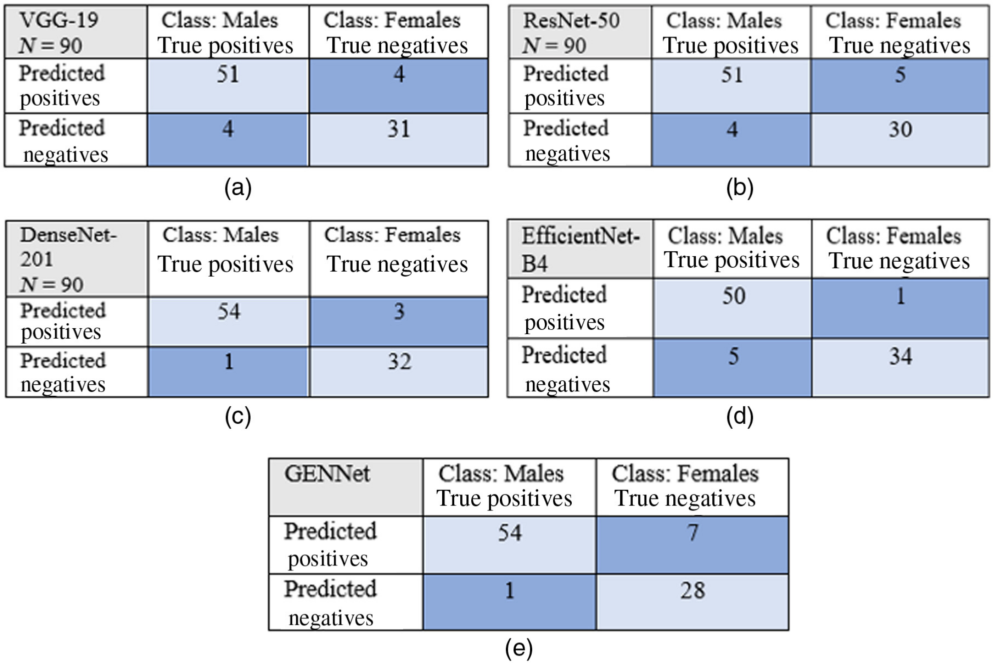
$$\text{accuracy}(\text{ACC}) = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}} \times 100, \tag{7}$$

where tp, fp, fn, and tn refer to true positive, false positive, false negative, and true negative, respectively. ACC in Eq. (7) means overall testing accuracy.

Figure 11 illustrates the calculated test accuracy along with total number of parameters chart of all the models. A confusion matrix for five of the best models is presented in Fig. 12 to better elaborate on the performance of each model on different genders.

By analyzing Fig. 11, we can observe that GENNet model performed significantly well among other low-parameter models by achieving total test accuracy of 91%, equal to the test accuracy of the VGG-19 model. However, VGG-19 has 138 million parameters, which is the highest number of parameters among all other models.

Figure 13 shows a number of failed predictions by the studied state-of-the-art models. The results display the model name along with the predicted output class.



VGG-19 $N = 90$	Class: Males True positives	Class: Females True negatives
Predicted positives	51	4
Predicted negatives	4	31

(a)

ResNet-50 $N = 90$	Class: Males True positives	Class: Females True negatives
Predicted positives	51	5
Predicted negatives	4	30

(b)

DenseNet-201 $N = 90$	Class: Males True positives	Class: Females True negatives
Predicted positives	54	3
Predicted negatives	1	32

(c)

EfficientNet-B4	Class: Males True positives	Class: Females True negatives
Predicted positives	50	1
Predicted negatives	5	34

(d)

GENNet	Class: Males True positives	Class: Females True negatives
Predicted positives	54	7
Predicted negatives	1	28

(e)

Fig. 12 Confusion matrix depicting the performance of (a) VGG-19; (b) ResNet-50; (c) DenseNet-201; (d) EfficientNet-B4; and (e) GENNet models.

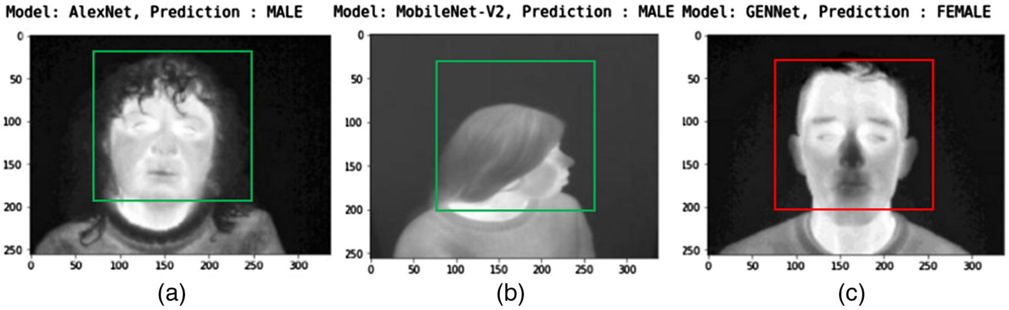

Model: AlexNet, Prediction : MALE
Model: MobileNet-V2, Prediction : MALE
Model: GENNet, Prediction : FEMALE
0
50
100
150
200
250
300
(a)
(b)
(c)

Fig. 13 Individual false prediction test case results: (a) AlexNet model: female gender misclassified as male gender; (b) MobileNet: female gender misclassified as male gender; and (c) GENNet: male gender misclassified as female gender.

In order to understand how effective, the models are for the custom classification task, eight different quantitative metrics are employed in addition to the accuracy metrics thus providing a detailed performance comparison of all the trained models. The additional metrics include sensitivity, specificity, precision, negative predictive value, false positive rate (FPR), false negative rate (FNR), Matthews correlation coefficient (MCC), and $F1$-score. Sensitivity, specificity, and precision are the conditional probabilities where sensitivity also termed as recall is defined as the probability of given positive example results in positive test, specificity is the probability of given negative example results in negative test, whereas precision provides what proportion of positive identifications was actually correct. The FPR is the proportion of negative cases incorrectly identified as positive cases in the data, whereas FNR also known as miss rate is the proportion of positive cases incorrectly identified as negative cases. $F1$-score describes the preciseness (such that how many instances it predicts correctly) and robustness (such that it does not miss a significant number of instances) of the classifier. MCC produces a more informative and reliable statistical score in evaluating binary classifications in addition to accuracy and $F1$-score. It produces a high score only if the trained classifier obtained good results in all the four confusion matrix categories including true positives, false negatives, true negatives, and false positives. The numerical results are presented in Table 5. The best and worst value per metric is highlighted in bold and italics.

5 Discussions

This section will discuss the overall performance of each model along with its individual training and inference time required compared to other models and individual parameters of each model. Table 6 presents the numerical values of this comparison.

- AlexNet model achieved the best inference time and sensitivity compared to the other models, but it has a low specificity and precision scores.
- EfficientNet-B4,[49] DenseNet-201, and GENNet model has achieved an optimal $F1$-score followed by VGG-19 and ResNet-50 architectures. Also EfficientNet-B4[49] achieved the highest testing accuracy of 93% and best MCC[61] scores, however, EfficientNet-B4 requires the highest training time.
- DenseNet-201 also proved to be one of the best models achieving the second best specificity and second lowest FPR. The total test accuracy of the model is 91%, however, it requires the highest inference time and relatively higher training time as compared to other models thus making it a computationally expensive model.
- The bigger architectures such as ResNet, DenseNet, and EfficientNet have good sensitivity and less FNR, however, the inference time required by these architectures is relatively high compared to other models.
- Although the proposed model GENNet has a high false-positive rate, but as a trade-off, it achieved the optimal test accuracy of 91% along with good sensitivity, $F1$ score, negative

Table 5 Different quantitative metrics. The best value per metric is highlighted in bold, and the worst value per metric is highlighted in italics.

Quantitative metrics comparison of all the models

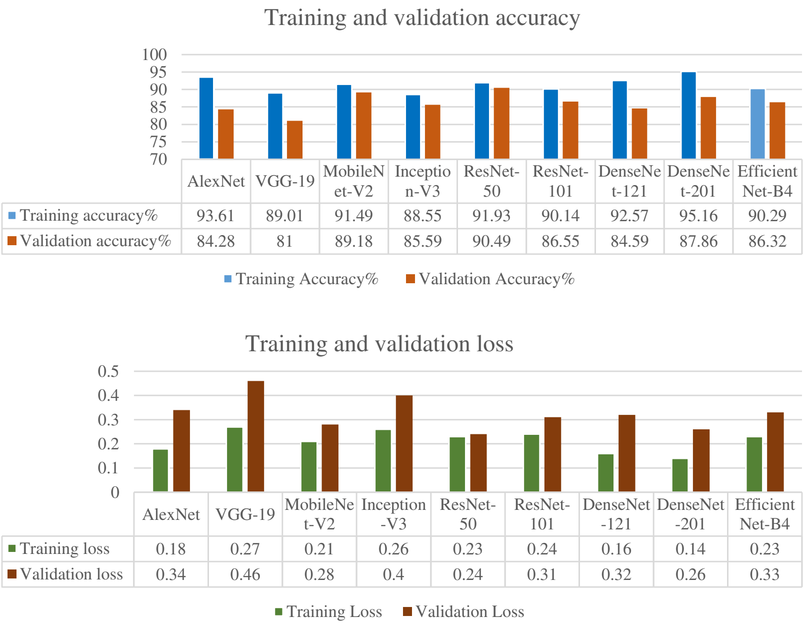
Models	Sensitivity	Specificity	Precision	Negative predictive value	FPR	FNR	*F*1-score	MCC
AlexNet	**0.98**	*0.54*	*0.77*	0.95	*0.45*	**0.02**	*0.86*	*0.61*
VGG-19	0.93	0.88	0.93	0.88	0.11	0.07	0.92	0.81
MobileNet-V2	*0.87*	0.86	0.90	*0.81*	0.14	*0.12*	0.89	0.72
Inception-V3	0.96	0.77	0.87	0.93	0.23	0.04	0.91	0.77
ResNet-50	0.93	0.85	0.91	0.88	0.14	0.07	0.92	0.78
ResNet-101	**0.98**	0.60	0.79	0.95	0.40	**0.02**	0.87	0.66
DenseNet-121	0.93	0.74	0.85	0.87	0.25	0.07	0.88	0.69
DenseNet-201	0.93	0.91	0.94	0.88	0.09	0.07	0.93	0.83
EfficientNet-B4	0.90	**0.97**	**0.98**	0.87	**0.03**	0.09	**0.94**	**0.86**
GENNet Model	**0.98**	0.80	0.89	**0.96**	0.20	**0.02**	0.93	0.82

Table 6 Comparison of total training and testing time required by all the models and individual model parameters

Models	Alex Net	VGG-19	Mobile Net-V2	Inception-V3	Res Net-50	Res Net-101	Dense Net-121	Dense Net-201	Efficient Net	GEN Net
Average training time required for each epoch (s)	2.66	12.19	4.55	6.2	6.4	10.3	8.3	11.33	15.13	3.1
Overall training time required (s)	266	1220	455	620	640	1030	830	1130	1513	310
Inference time required for complete test data (s)	3.6	13.2	4.1	8.3	7.2	11.2	7.4	9.3	7.2	3.6
Parameters (million)	62.3	138	2.2	24	26	43	7.2	18.6	19 M	16.8

predictive value, and lowest FNR when compared to other low or nearly equivalent parameter models. In addition to this, the model requires the least inference time like AlexNet.

- By analyzing the low-specificity value of all the models except EfficientNet-B4 compared to the sensitivity metric as shown in Table 7, it can be concluded that low can be overcome by using a significant amount of thermal training data to better generalize the capabilities of DNN.
- Moreover, currently, the main focus is on gender classification for in-cabin driver monitoring systems using thermal facial features. The current technique can be expanded to face recognition and obtaining other biometrics information in random outdoor environmental conditions. For instance, in law enforcement applications[62] this system can be made more effective by capturing data through CCTV recordings. The recorded data can be used for training and thus performing multi-frame detection and classification tasks such as hat and mask detection, and then subsequently classifying the person's gender. This can be achieved by training advanced deep learning algorithms[63,64] such as human body instance segmentation and recognition.



	AlexNet	VGG-19	MobileNet-V2	Inception-V3	ResNet-50	ResNet-101	DenseNet-121	DenseNet-201	EfficientNet-B4
Training accuracy%	93.61	89.01	91.49	88.55	91.93	90.14	92.57	95.16	90.29
Validation accuracy%	84.28	81	89.18	85.59	90.49	86.55	84.59	87.86	86.32

	AlexNet	VGG-19	MobileNet-V2	Inception-V3	ResNet-50	ResNet-101	DenseNet-121	DenseNet-201	EfficientNet-B4
Training loss	0.18	0.27	0.21	0.26	0.23	0.24	0.16	0.14	0.23
Validation loss	0.34	0.46	0.28	0.4	0.24	0.31	0.32	0.26	0.33

Fig. 14 Accuracy and loss charts of all the networks trained using freezed layer configuration.

6 Ablation Study

This section shows an ablation study by analyzing the results of the nine state-of-the-art deep learning networks by freezing the network layers as discussed in Sec. 3.1. Figure 14 presents the overall performance of all the pretrained architectures initially trained on Casia dataset[42] and fine-tuned on thermal facial images from Tufts dataset.[10–12] The networks were trained using both SGD and Adam optimizer, and the best training and validation results in the case of each model were selected. It is important to mention that during the training phase the data are divided subject-wise and all the eight poses of each particular subject are used for training and validation purposes, respectively. This is done to avoid bias and to do optimal inductive learning. Figure 14 presents the training and validation accuracy and loss chart of all the pretrained models.

Among all the models ResNet-50 architecture scores highest with the validation accuracy of 90.49% followed by MobileNet-V2 with a validation accuracy of 89.18% using the SGD optimizer. However, AlexNet, VGG, and EfficientNet architectures do not perform well as compared to other models thus getting the lower validation accuracy and higher loss values. However, it was not possible to achieve an optimal training outcome as most of the models have accuracy levels below 95% with freeze layer configuration. By analyzing the accuracy and loss charts in Fig. 14, it is clear that during the finetuning process of all the pretrained models DenseNet-201[48] and AlexNet achieves the highest training accuracies of 95.16% (using SGD optimizer) and 93.61% (using Adam optimizer) with the lowest training losses of 0.14 and 0.18, respectively. MobileNet-V2[47] architecture achieved the best validation accuracy of 89.18% with a validation loss of 0.28 (using SGD optimizer). However, it achieved a lower training accuracy of 90.32% with validation accuracy of 90.16% when the model was trained using Adam optimizer. The DenseNet-201 model scored second best with a validation accuracy of nearly 88% (using SGD optimizer). The VGG-19 architecture was unable to achieve good accuracy scores compared to the other pretrained models with overall validation accuracy of only 81% and the highest validation loss of 0.46.

7 Conclusions and Future Work

In the proposed study, we have proposed a new CNN architecture GENNet for autonomous gender classification using thermal images. Initially, all the models including pretrained models as well as newly proposed GENNet models are trained on a large-scale human facial structures, which eventually help us to fine-tune the model on smaller thermal facial data more robustly. In order to achieve optimal training accuracy and less error rate, all the networks are trained using two different state-of-the-art optimizers including SGD and Adam optimizers and picked the best results in the case of each model. The trained models are cross-validated using two new thermal datasets including the public as well as the locally gathered dataset. The EfficientNet-B4 model achieved the highest training accuracy of 93% followed by the DenseNet-201, and the proposed network has achieved an overall testing accuracy of 92% and 91%. However, GENNet architecture is good for a compute-constrained thermal gender classification use-case as it performs significantly better than other low-parameter models.

For future work, we can work on the grouping of different datasets and fusions of features that can eventually push toward the horizon for the advancement of deep learning. In the same way, we can use techniques to generate new data from the existing data such as smart augmentation techniques, GANs, and last but not least generating synthetic data that can aid us in increasing the accuracy levels and reducing the overfitting of a target network. Moreover, multi-scale convolutional neural networks can be designed for performing more than one human biometrics task such as face recognition, age estimation, and emotion recognition using thermal data. For example, face recognition using thermal imaging can be performed using blood perfusion data by extracting blood vessels patterns, which are unique in all human beings. Similarly, emotion recognition can be performed by learning specific thermal patterns in human faces while recording different emotions.

Appendix A

Table 7 shows the complete layer-wise architectural details of the newly proposed GENNet model for task-specific thermal gender classification.

Table 7 Layer wise architecture of GENNet. Output shape is shown in brackets along with kernel size, no of stride, padding, and number of network parameters

Block-1	Block-2	Block-3	Block-4
Conv 2D-1 [16, 16, 250, 250] Kernel size = 3 Stride = 1 Padding = 1 No of param = 448	Conv 2D-5 [16, 32, 125, 125] Kernel size = 3 Stride = 1 Padding = 1 No of param = 4,640	Conv 2D-9 [32, 64, 62, 62] Kernel size = 3 Stride = 1 Padding = 1 No of param = 18,496	FC-1/linear-13 [65536, 256] No of param = 16,777,472
ReLU-2 [16, 16, 250, 250]	ReLU-6 [16, 32, 125, 125]	ReLU-10 [32, 64, 62, 62]	ReLU-14
MaxPool 2D-3 [16, 16, 125, 125] Kernel size = 2 Stride = 2	MaxPool 2D-7 [16, 32, 62, 62] Kernel size = 2 Stride = 2	MaxPool 2D-11 [32, 64, 32, 32] Kernel size = 2 Stride = 2 Padding = 1	Dropout (0.5)-15
Dropout (0.5)-4 [16, 16, 125, 125]	Dropout (0.5)-8 [16, 32, 62, 62]	Dropout (0.3)-12 [32, 64, 32, 32]	FC-2/linear [256, 1] Total no of param = 16,801,570

Appendix B

During the experimental work, when training the GENNet model from scratch using only thermal dataset, we were unable to achieve precise training and validation accuracy with greater loss values, which eventually results in low testing accuracy. The experiments were carried using different optimizers including adaptive learning rate optimization Adam[56] as well as SGD,[55] but the same results were observed. The experimental results are demonstrated in Fig. 15.

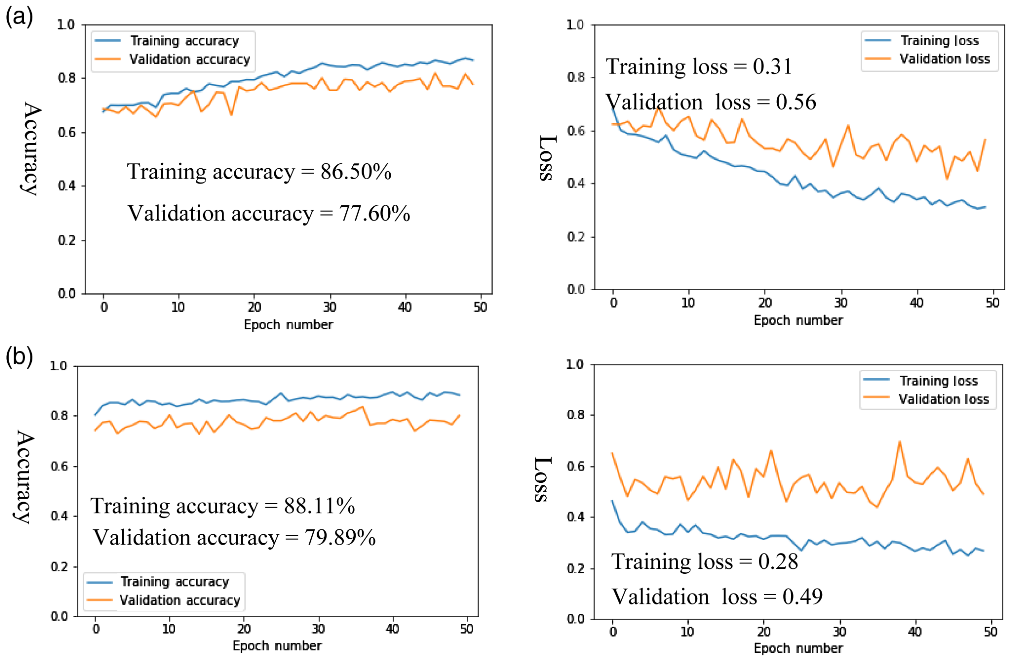


Fig. 15 Training GENNet accuracies and loss graph using only thermal data: (a) training and validation accuracy and loss graph using Adam optimizer and (b) training and validation accuracy and loss using SGD optimizer.

Acknowledgments

This thermal gender classification system using the public as well locally gathered dataset acquired using prototype thermal camera with measured accuracies of state-of-the-art models is part of the project that has received funding from the ECSEL Joint Undertaking (JU) under Grant Agreement No 826131. The JU receives support from the European Union's Horizon 2020 research and innovation program and the national funding from France, Germany, Ireland (Enterprise Ireland International Research Fund), and Italy. The authors would like to acknowledge Joesph Lamley for providing his support on how to regularize and generalize the new DNN architecture with smaller datasets, Xperi Ireland team, Chris Dainty, and Quentin Noir from Lynred France for giving their feedback. Moreover, the authors would like to acknowledge Tufts University for the contributors of the Tufts dataset and Carl dataset for providing the image resources to carry out this research work. Authors have no relevant financial interests in the manuscript and no other potential conflicts of interest to disclose. For the proposed study informed consent was obtained from all the five subjects to publish their thermal facial data.

References

1. Heliaus European Union Project, https://www.heliaus.eu/ (accessed 20 January 2020).
2. Y. Abdelrahman et al., "Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **1**(3), 1–20 (2017).

3. A. Raahul et al., "Voice based gender classification using machine learning," *IOP Conf. Series: Mat. Sci. Eng.* **263**(4), 042083 (2017).
4. A. Abdelwhab and S. Viriri, "A survey on soft biometrics for human identification," in *Machine Learning and Biometrics*, J. Yang et al., Eds., p. 37 (2018).
5. S. Karjalainen, "Thermal comfort and gender: a literature review," *Indoor Air* **22**(2), 96–109 (2012).
6. V. Espinosa-Duró et al., "A criterion for analysis of different sensor combinations with an application to face biometrics," *Cognit. Comput.* **2**(3), 135–141 (2010).
7. D. A. Lewis, E. Kamon, and J. L. Hodgson, "Physiological differences between genders implications for sports conditioning," *Sports Med.* **3**(5), 357–369 (1986).
8. J. Christensen, M. Væth, and A. Wenzel, "Thermographic imaging of facial skin—gender differences and temperature changes over time in healthy subjects," *Dentomaxillofacial Radiol.* **41**(8), 662–667 (2012).
9. J. D. Bronzino and D. R. Peterson, *Biomedical Signals, Imaging, and Informatics*, CRC Press, Boca Raton, Florida (2014).
10. K. Panetta et al., "The tufts face database," http://tdface.ece.tufts.edu/ (accessed on 29 October 2019).
11. K. Panetta et al., "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 509–520 (2020).
12. K. M. S. Kamath et al., "TERNet: a deep learning approach for thermal face emotion recognition," *Proc. SPIE* **10993**, 1099309 (2019).
13. V. Espinosa-Duró, M. Faundez-Zanuy, and J. Mekyska, "A new face database simultaneously acquired in visible, near-infrared and thermal spectrums," *Cognit. Comput.* **5**(1), 119–135 (2013).
14. E. Makinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(3), 541–547 (2008).
15. D. A. Reid et al., "Soft biometrics for surveillance: an overview," in *Handbook of Statistics*, C. R. Rao and V. Govindaraju, Vol. **31**, pp. 327–352, Elsevier, North Holland (2013).
16. G. Guo and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image Vision Comput.* **32**(10), 761–770 (2014).
17. A. J. O'Toole et al., "Sex classification is better with three-dimensional head structure than with image intensity information," *Perception* **26**, 75–84 (1997).
18. B. Moghaddam and M.-H. Yang, "Learning gender with support faces," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(5), 707–711 (2002).
19. Y. Elmir, Z. Elberrichi, and R. Adjoudj, "Support vector machine based fingerprint identification," in *Conférence nationale sur l'informatique et les Technologies de l'Information et de la Communication*, Vol. **2012** (2012).
20. S. Baluja and H. A. Rowley, "Boosting sex identification performance," *Int. J. Comput. Vision* **71**(1), 111–119 (2007).
21. M. Toews and T. Arbel, "Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(9), 1567–1581 (2009).
22. I. Ullah et al., "Gender recognition from face images with local wld descriptor," in *19th Int. Conf. Syst., Signals and Image Process.*, IEEE (2012).
23. J. Chen et al., "WLD: a robust local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1705–1720 (2010).
24. P. J. Phillips et al., "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vision Comput.* **16**(5), 295–306 (1998).
25. C. Perez et al., "Gender classification from face images using mutual information and feature fusion," *Int. J. Optomechatron.* **6**(1), 92–119 (2012).
26. K. S. Arun and K. S. A. Rarath, "Machine learning approach for fingerprint based gender identification," in *Proc. IEEE Conf. Recent Adv. Intell. Comput. Syst.*, Trivandrum, India, pp. 163–16 (2011).
27. C. Chen and A. Ross, "Evaluation of gender classification methods on thermal and near-infrared face images," in *Int. Joint Conf. Biom.*, IEEE (2011).

28. D. T. Nguyen and K. R. Park, "Body-based gender recognition using images from visible and thermal cameras," *Sensors* **16**(2), 156 (2016).
29. L. Xiao et al., "Combining HWEBING and HOG-MLBP features for pedestrian detection," *J. Eng.* **2018**(16), 1421–1426 (2018).
30. M. Abouelenien et al., "Multimodal gender detection," in *Proc. 19th ACM Int. Conf. Multimodal Interaction* (2017).
31. H. Malik et al., "Applications of artificial intelligence techniques in engineering," in *SIGMA*, Vol. **698** (2018).
32. A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," arXiv:1605.07678 (2016).
33. N. Dwivedi and D. K. Singh, "Review of deep learning techniques for gender classification in images," in *Harmony Search and Nature Inspired Optimization Algorithms*, N. Yadav et al., Eds., Vol. **741**, pp. 327–352, Springer, Singapore (2019).
34. Mivia Lab University of Salerno, "Gender-FERET dataset," http://mivia.unisa.it/database/gender-feret.zip (accessed 30 June 2020).
35. G. Ozbulak, Y. Aytar, and H. K. Ekenel, "How transferable are CNN-based features for age and gender classification?" in *Int. Conf. Biom. Special Interest Group*, IEEE (2016).
36. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent. (ICLR)*, San Diego, California (2015).
37. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.* (2012).
38. E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Secur.* special issue on Facial Biometrics in the Wild **9**(12), 2170–2179 (2014).
39. A. Manyala et al., "CNN-based gender classification in near-infrared periocular images," *Pattern Anal. Appl.* **22**(4), 1493–1504 (2019).
40. O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *Proc. British Machine Vision Conf. (BMVC)*, pp. 1–12, BMVA Press (2015).
41. N. R. Baek et al., "Multimodal camera-based gender recognition using human-body image with two-step reconstruction network," *IEEE Access* **7**, 104025–104044 (2019).
42. D. Yi et al., "Learning face representation from scratch," arXiv:1411.7923 (2014).
43. FLIR, "FLIR Vuo Pro thermal camera," https://www.flir.com/products/vue-pro/ (accessed 14 October 2019).
44. J. Deng et al., "Imagenet: a large-scale hierarchical image database," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE (2009).
45. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2016).
46. C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2016).
47. M. Sandler et al., "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2018).
48. G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2017).
49. M. Tan and Q. V. Le. "Efficientnet: rethinking model scaling for convolutional neural networks," *Proceedings of the 36th International Conference on Machine Learning*, Vol. **97**, pp. 6105–6114 (2019).
50. S. Mallick, "Image classification using transfer learning in Pytorch," https://www.learnopencv.com/image-classification-using-transfer-learning-in-pytorch/ (accessed 10 January 2020).
51. A. Khan et al., "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.* **53**, 5455–5516 (2020).
52. V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.* (2010).
53. P. Smith and C. Chen, "Transfer learning with deep CNNs for gender recognition and age estimation," in *IEEE Int. Conf. Big Data*, IEEE, Seattle, Washington, pp. 2564–2571 (2018).
54. "Pytorch deep learning framework," https://pytorch.org/ (accessed 14 October 2019).

55. L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, Physica-Verlag HD, pp. 177–186 (2010).
56. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
57. Lynred France, "Heliaus project coordinator and consortium partner," https://www.lynred.com/ (accessed 27 January 2020).
58. "Camera optics measurements," https://www.edmundoptics.eu/knowledge-center/application-notes/imaging/understanding-focal-length-and-field-of-view/ (accessed 15 February 2020).
59. A. Tempelhahn et al., "Shutter-less calibration of uncooled infrared cameras," *J. Sens. Sens. Syst.* **5**(1), 9 (2016).
60. M. Stojanovi et al., "Understanding sensitivity, specificity, and predictive values," *Vojnosanit Pregl* **71**(11), 1062–1065 (2014).
61. B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta* **405**(2), 442–451 (1975).
62. M. Zabłocki et al., "Intelligent video surveillance systems for public spaces—a survey," *J. Theor. Appl. Comput. Sci.* **8**(4), 13–27 (2014).
63. K. He et al., "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, pp. 2961–2969 (2017).
64. M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," arXiv:1701.04862 (2017).

Muhammad Ali Farooq received his BE degree in electronics engineering from IQRA University in 2012 and his MS degree in electrical control engineering from the National University of Sciences and Technology in 2017. He is a PhD researcher at the National University of Ireland Galway. His research interests include machine vision, computer vision, smart embedded systems, and sensor fusion. He has won the prestigious H2020 European Union (EU) scholarship and currently working on safe autonomous driving systems under the HELIAUS EU project.

Hossein Javidnia received his PhD in electronic engineering from the National University of Ireland Galway focused on depth perception and 3D reconstruction. He is a research fellow at ADAPT Centre, Trinity College, Dublin, Ireland, and a committee member at the National Standards Authority of Ireland working on the development of a national AI strategy in Ireland. He is currently researching offline augmented reality and generative models.

Peter Corcoran is the editor-in-chief of the IEEE Consumer Electronics Magazine and a professor with a personal chair at the College of Engineering and Informatics of NUI Galway. In addition to his academic career, he is also an occasional entrepreneur, industry consultant, and compulsive inventor. His research interests include biometrics, cryptography, computational imaging, and consumer electronics.

Appendix D

C3I Thermal Automotive Dataset

Authors' Contribution to [17]

Contribution Criteria	*Contribution Percentage*
Data Acquisition Setup	MAF: 60%, CR: 40%
Dataset Collection	MAF: 60%, PC: 40%
Dataset Preparation	MAF: 100%

C3I THERMAL AUTOMOTIVE DATASET

Citation Author(s): Muhammad Ali Farooq (https://orcid.org/0000-0003-4116-8021) *(National University of Ireland Galway (NUIG))*
Waseem Shariff (https://orcid.org/0000-0001-7298-9389) *(National University of Ireland Galway (NUIG))*
Faisal Khan (https://orcid.org/0000-0002-8391-6203) *(National University of Ireland Galway (NUIG))*
Peter Corcoran (https://orcid.org/0000-0003-1670-4793) *(National University of Ireland Galway (NUIG))*
Cosmin Rotariu *(Xperi Corporation)*

Submitted by: Muhammad Ali Farooq

Last updated: Sat, 03/26/2022 - 12:16

DOI: 10.21227/jf21-rt22

Data Format: *.avi; *.png; *.csv; *.zip

Links: C3I (Center for Computational, Cognitive and Connected Imaging) (https://nuigalway.ie/c3i/)
Heliaus (tHErmaL vIsion AUgmented awarenesS) (https://www.heliaus.eu/)

Categories:	Artificial Intelligence Digital signal processing Machine Learning Image Processing Computational Intelligence Sensors
Keywords:	Thermal imaging, ADAS, LWIR, Object detection dataset, Autonomous vehicle

0 ratings - Please login to submit your rating.

ACCESS DATASET CITE SHARE/EMBED

ABSTRACT

The C3I Thermal Automotive Dataset provides > 35,000 distinct frames along with annotated thermal frames for the development of smart thermal perception system/ object detection system that will enable the automotive industry and researchers to develop safer and more efficient ADAS and self-driving car systems. The overall dataset is acquired, processed, and open-sourced in challenging weather and environmental scenarios. The dataset is recorded from a lost-cost yet effective 640x480 uncooled LWIR thermal camera. The dataset is gathered by mounting the camera stand-alone and on an electric vehicle to minimize mechanical vibrations.

Instructions:
This Dataset is Published by NUIG under HELIAUS EU Project.
All researchers/professionals need to follow the instructions below to access the datasets.

• If you are using this dataset it is requested to fill the google dataset dissemination form (Link: https://docs.google.com/forms/d/e/1FAIpQLSfi8K8DE4a6kBtVD3GyCXvMCcTrvSCm... (https://docs.google.com/forms/d/e/1FAIpQLSfi8K8DE4a6kBtVD3GyCXvMCcTrvSCmvq9Y9LqHb0FTm1E8Pg/viewform)). Please use institutional/ Company email address(es). Commercial emails such as Gmail are not allowed.
• Annotation for bounding boxes is provided to train YOLO/ YOLO-v5 based detectors.
• Also please cite this dataset along with the following papers.

1. M. A. Farooq, P. Corcoran, C. Rotariu and W. Shariff, "Object Detection in Thermal Spectrum for Advanced Driver-Assistance Systems (ADAS)," in IEEE Access, vol. 9, pp. 156465-156481, 2021, doi: 10.1109/ACCESS.2021.3129150.
Link: https://ieeexplore.ieee.org/document/9618926 (https://ieeexplore.ieee.org/document/9618926)

2. M. A. Farooq, W. Shariff and P. Corcoran, "Evaluation of Thermal Imaging on Embedded GPU Platforms for Application in Vehicular Assistance Systems," in IEEE Transactions on Intelligent Vehicles, doi: 10.1109/TIV.2022.3158094.
Link: https://ieeexplore.ieee.org/document/9732195 (https://ieeexplore.ieee.org/document/9732195)

Thank you

Best Regards
Muhammad Ali Farooq
PhD Researcher
Email: m.farooq3@nuigalway.ie (mailto:m.farooq3@nuigalway.ie) ; mail.frq@gmail.com (mailto:mail.frq@gmail.com)
Cell no: +3530899556961
NUIG-Ireland

Funding Agency: ECSEL Joint Undertaking (JU) under grant agreement No 826131

COMMENTS

()

Submitted by Muhammad Ali Farooq

Submitted by Muhammad Ali Farooq (/authors/muhammad-ali-farooq) on Sat, 03/26/2022 - 12:19

Log in (/saml_login?destination=node/9170%23comment-form) to post comments

DATASET FILES

- Thermal Frames 640x480.7z (482.66 MB)
- Video Sets.7z (263.58 MB)
- Annotated Data.7z (67.95 MB)
- data-annotation-tool.7z (6.22 MB)

LOGIN TO ACCESS DATASET FILES

DOCUMENTATION

Complete Dataset Details - NUIG -2022.pdf (402.46 KB)

QUESTIONS?

✉ Login to Send Author a Private Message

⚠ Report a problem with this Dataset

Appendix E

Object Detection in Thermal Spectrum for Advanced Driver-Assistance Systems (ADAS)

Authors' Contribution to [18]

Contribution Criteria	*Contribution Percentage*
Research Hypothesis	MAF: 80%, PC: 20%
Experiments and Implementation	MAF: 100%
Background	MAF: 100%
Manuscript Preparation	MAF: 60%, WS: 20% PC: 15%, CR: 5%

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 2, 2021, accepted November 15, 2021, date of publication November 17, 2021,
date of current version December 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129150

Object Detection in Thermal Spectrum for Advanced Driver-Assistance Systems (ADAS)

MUHAMMAD ALI FAROOQ[1], PETER CORCORAN[1], (Fellow, IEEE), COSMIN ROTARIU[2], AND WASEEM SHARIFF[1]
[1]School of Engineering, National University of Ireland Galway, Galway, H91 TK33 Ireland
[2]Xperi Corporation, Galway, H91 V0TX Ireland

Corresponding author: Muhammad Ali Farooq (m.farooq3@nuigalway.ie)

This research work is carried under Heliaus (thermal vision augmented awareness) EU project (website: www.heliaus.eu). The project has received funding from the Electronic Components and Systems for European Leadership (ECSEL) Joint Undertaking (JU) under grant agreement no 826131. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Germany, Ireland, Italy.

ABSTRACT AI-based smart thermal perception systems can cater to the limitations of conventional imaging sensors by providing a more reliable data source in low-lighting conditions and adverse weather conditions. This research evaluates and modifies the state-of-the-art object detection and classifier framework for thermal vision with seven key object classes in order to provide superior thermal sensing and scene understanding input for advanced driver-assistance systems (ADAS). The networks are trained on public datasets and is validated on test data with three different test approaches which include test-time augmentation, test-time with no augmentation, and test-time with model ensembling. Additionally, a new model ensemble-based inference engine is proposed, and its efficacy is tested on locally gathered novel test data comprising of 20K thermal frames captured with an uncooled LWIR prototype thermal camera in challenging weather and environmental scenarios. The performance analysis of trained models is investigated by computing precision, recall, and mean average precision scores (mAP). Furthermore, the smaller network variant of thermal-YOLO architecture is optimized using TensorRT inference accelerator, which is then deployed on GPU and resource-constrained edge hardware Nvidia Jetson Nano. This is implemented to explicitly reduce the inference time on GPU as well as on Nvidia Jetson Nano to evaluate the feasibility for added real-time onboard installations.

INDEX TERMS Thermal-infrared, object detection, advanced driver-assistance systems, deep learning, edge computing.

I. INTRODUCTION

Thermal cameras can be used for object detection in both day and night-time environmental conditions [1]. Since it is invariant to illumination changes, occlusions, and shadows it provides improved situational awareness. Moreover, by integrating with AI-based imaging pipelines we can design intelligent thermal perception systems to detect multiple objects of different classes. Such systems can be beneficial for advanced driver assistance systems (ADAS) & environment monitoring methods. Vehicular perception systems has become an emerging consumer technology application and the evolution of this technology over time aims to provide extended safety benefits and reliable means of transportation. Various key technologies are directly associated with intelligent vehicular systems which includes, sensor fusion for real-time data logging, and object/ obstacle detection and tracking system using machine learning algorithms. This will empower the drivers to monitor the external environment, detecting external objects, and predict events that the driver needs to be aware of thus providing a deeper understanding of the entire road surroundings.

There is a range of sensors commonly used for designing smart perception systems for automative sensor suite such as lidar and radar. Practical systems often leverage both visible imaging solutions along with the array of hardware sensors. However, visible imaging has some limitations. For instance, the RGB camera operates inadequately in unfavorable illumination conditions such as low lighting, sun glare, and glare from the headlight beam. Moreover, typical automative sensors (radar and lidar) leverage some drawbacks in computer vision applications as discussed in [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Khin Wee Lai.

Recent developments in microbolometer technology have led to lower costs for uncooled thermal imaging sensors. These sensors in the automotive suite can complement or even be integrated with existing technology, offering a particular advantage that as they sense the thermal emissivity of objects, and they can operate independently of lighting conditions, therefore, making it a more consistent solution for enhanced environmental perception systems [3].

Object detection plays a key role in designing intelligent perception systems. However, the robustness of object detection algorithms on thermal data has yet not achieved vigorous results and is still a challenging research area, in the field of computer vision [4]. It is due to two important factors which include lack of availability of large-scale thermal datasets as compared to visible datasets [3] and secondly most of the established DNN architectures are pre-trained on visible datasets [5], [6] thus making it a stimulating task to appropriately converge on thermal data features which is an essential requirement for robust training of deep learning models. Moreover, some of the constraints in publicly available datasets also lead towards the challenge in optimum training of dense models on thermal imaging. Such that some of the publicly available thermal data consist of video sequences however with little variability in the scene, i.e., weather conditions, light conditions, and person heat radiation. This drawback reduces the generalization of object detectors. Furthermore, some of the thermal datasets available for roadside object detection are mainly captured from the frontal view [7]. To overcome such complexities four different thermal public datasets are employed for including sufficient data diversity and robust training of deep learning models as shown in table 1. These datasets are captured in different environmental and weather conditions, varying object distance from the camera, and different view angles thus making it beneficial for optimal training and proper generalization of YOLO object detection algorithm.

In this study, we have focused on thermal object detection using state-of-the-art YOLO-v5 [8], [9] end-to-end deep learning framework as it can play a key element in the successful implementation and deployment of object detection and classification at thermal wavelengths. The main goal is to achieve robust inference results by doing optimal training and fine-tuning of CNN architectures using two different optimizers (SGD and ADAM) and selecting an appropriate set of network hyperparameters. The efficacy of trained networks is validated and computed using various accuracy metrics by running the inference test on complex test data accumulated from unseen public data and locally acquired novel test data. The overall validation tests are performed by incorporating three different test approaches which include test-time with no augmentation (TTNA), test-time augmentation (TTA), and test-time with model ensembling (TTME) for improved test accuracy.

TABLE 1. Publicly available thermal datasets attributes.

Datasets	Weather Conditions and viewpoint	Environment	No of Frames	Objects	Camera specifications and image resolution
CVC-09 [34]	Daytime and nighttime nearside	roadside	11K	Person Cars Poles Bicycle Bus Bikes	640 x 480
FLIR ADAS [36]	Daytime and nighttime with Sun and cloudy condition Nearside	roadside	14K	Person Cars Poles Bus Dog Bicycles	FLIR Tau2 LWIR camera 640 x 512
OSU Thermal [16]	Daytime with haze, fair, light rain, and partially cloudy weather conditions far top	University campus surveillance environment. Data was gathered by mounting the camera on the rooftop of an 8-story building	284	Person Cars Poles	Raytheon 300D sensor core with 75 mm lens 30 Hz 320 x 240
Four sets from KAIST Multispectral dataset [32]	Set 00 Daytime Nearside	campus	17,498	Person Cars Poles Bicycles Bus	20 Hz 640 x 480
	Set 01 Daytime Nearside	roadside	8,035		
	Set 04 Nighttime Nearside	roadside	7.2K		
	Set 05 Nighttime nearside	downtown	2.9K		

A. MAIN CONTRIBUTIONS OF THIS PAPER THE CORE CONTRIBUTIONS OF THIS RESEARCH WORK INCLUDE

- Adaptation and validation of a state-of-the-art object detection/ classification framework for designing smart thermal perception system with seven distinct classes including stationary as well as moving objects. Moreover, a new model ensemble-based inference engine is proposed using the combination of two best-trained models to further improve the accuracy metrics on test data.
- A novel test dataset is captured using an uncooled LWIR thermal camera developed under Heliaus EU project [10] in different environments and

weather conditions. A total of 20,000 thermal frames have been acquired and selected for this study consisting of seven different class objects.

- Evaluating a neural framework with a range of model sizes to determine its suitability for porting to a resource-constrained embedded edge platform (Nvidia Jetson). Thus, to study its feasibility for further automotive on-board-computer (OBC) installations [11].

II. BACKGROUND/RELATED WORK

Common practices in ADAS architecture have been established over the years. Most of these systems divide the task of safe and advanced driving into subcategories and employ an array of sensors and algorithms on various hardware modules for diversified tasks. Machine learning and specifically deep learning models [12] have become dominant in many of these tasks among which object detection is one of them. This section will mainly focus on the published studies exploring state-of-the-art object detection methods, and the reported results that investigate the area of object detection in the thermal spectrum. It includes various object classes such as pedestrian and vehicle detection.

A. OBJECT DETECTION IN THERMAL SPECTRUM USING MULTIMODAL AND DEEP LEARNING METHODS

The first phase explores multimodal machine learning methods that mainly rely on manually extracted feature vectors which are then fed to different types of classifiers and detectors for performing object detection in thermal spectrum either offline or in real-time. Olmeda *et al.* [13], proposed pedestrian detection in FIR images by using a new feature descriptor, the histograms of oriented phase energy (HOPE), and an adaptation of the latent variable based on the support vector machine (SVM). The authors concluded that histogram-based features perform exceptionally well as compared to other Linear binary patterns (LBP) and Principal Component Analysis (PCA). Besbes *et al.* [14] propose a pipeline for pedestrian detection in thermal images by using a hierarchical codebook of Speeded Up Robust Features (SURF) in the head region, taking advantage of the brightness of this area inside the regions of interest (ROIs). The reported experimental results show improved accuracy as compared to Haar-like Adaboost-cascade, linear SVM, and MultiFtr pedestrian detectors, trained on the FIR images. Coming towards more recent studies Lahmayed *et al.* [15] presented a method based on three different feature extractors. It includes multi-threshold and Histogram of Oriented Gradients (HOG) and Histograms of Oriented Optical Flow (HOOF) colour features combined with an SVM using both thermal infrared and visible light images. The authors validated their algorithm on three different datasets i.e., OSU colour thermal dataset [16], video analytic dataset, and LITIV dataset [17]. However, the main drawback of conventional machine learning classifiers such that SVM [18] cannot perform well with big datasets, and noisy data such that target classes are overlapping with each other.

The second phase explores deep learning and most specifically convolution neural networks (CNN) which have become an emerging trend for building intelligent imaging pipelines. This is due to the fact that end-to-end CNN models have proved their strengths in various computer vision applications by achieving robust and precision accuracy as compared to multimodal conventional machine learning methods. There are various state-of-the-art published deep learning-based object detection frameworks which include YOLO [19], Single Shot MultiBox Detector (SSD) [20], R-CNN [21], Fast R-CNN [22], and Mask R-CNN [23]. However, all these frameworks are built, trained, and tested on visible data. In this study, the prime focus is to explore the robustness of the object detector i.e., YOLO in thermal infrared spectrums. Various published studies [24]–[28] can be found where deep learning is employed for object detection in thermal images. In these studies, authors have used thermal data for object detection i.e., pedestrian detection in differing illumination conditions. The system works by extracting the feature maps from multispectral images. In the next step, these feature maps are fed to state-of-art object detectors which include faster-RCNN [29] and YOLO [19]. Recently Authors in [30] have used a YOLO detector for automatic human detection for surveillance application. The results concluded that due to the vast variation between visual and thermal data, the original YOLO model [19] has not achieved satisfactory outcomes by scoring the average precision (AP) of just 23% for single class person detection task in different weather conditions. Herrmann *et al.* [31] tested the Single Shot Detector (SSD) object detector by applying different pre-processing techniques to assess the performance of the detector on thermal data. They used KAIST [32] dataset for performance evaluation. The authors also worked with Maximally Stable Extremal Regions (MSERs) and later on classify the detected proposals by using CNN. The approach was tested on the OSU thermal pedestrian [14], OSU colour thermal [16], and Terravic motion IR [33] datasets. Recently, Huda *et al.* [7] used a YOLO object detector for person detection in the thermal spectrum. The authors had created their outdoor thermal dataset for transfer learning the YOLO-v3. The trained models were tested on three different public datasets which include CVC-09 [34], OSU-Thermal [16], and BU-TIV-atrium [35]. Similarly, in another recent study by Farzeen *et al.* [3] authors have explored domain adaption through style transfer methodology. These authors have used GAN architectures and cross-domain models on thermal and visible spectrum images. In the next stage, the style consistency approach is used for object detection by using two different public datasets, FLIR ADAS [36] and Kaist Multi-spectral dataset [32]. The authors established that adapting the low-level features from source domain to target domain using domain adaptation increases the mean average precision by approximately 10%. As per our finding, no published studies can be listed where authors have investigated the real-time feasibility of object detection algorithms in the thermal spectrum on edge devices for ADAS applications.

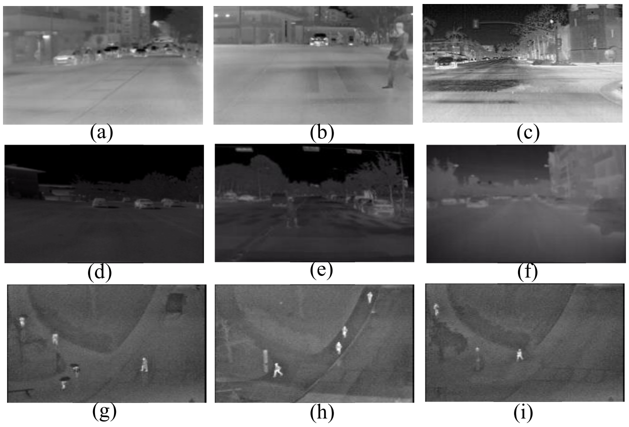


FIGURE 1. Sample images under varying environmental conditions from four different public datasets (a) CVC-09 dataset: frame acquired in day time road environment, (b) CVC-09: frame acquired in night-time road environment, (c) FLIR ADAS dataset: frame acquired in cloudy weather and road environment, (d) KAIST dataset: frame acquired in day time campus environment, (e) KAIST dataset: frame acquired in day time road environment, (f) KAIST dataset: frame acquired in night-time road environment, (g) OSU-thermal dataset: frame acquired in day-time with light rain in Ohio State University campus, (h) OSU-thermal dataset: frame acquired in day-time with partly cloudy weather in the campus environment and (i) OSU-thermal dataset: frame acquired in day-time with haze/ dusty weather conditions in the university campus environment.

III. PROPOSED METHODOLOGY

This section mainly focuses on the proposed methodology for robust training of state-of-the-art YOLO-v5 framework [8], [9] for out-cabin object detection in the thermal spectrum. Yolo was first introduced by Redmon *et al.* [19] for real-time object detection. YOLO is considered one of the fastest and finest deep learning algorithms for object detection in images, videos, and real-time camera streams.

The algorithm utilizes regression techniques thus training the whole image at once to optimize the overall performance. Moreover, it detects the class objects with their probabilities scores at the same time without requiring region proposals. YOLO-v5 is natively implanted in PyTorch whereas all prior models in the YOLO family leverage Darknet deep learning framework [37]. The networks are trained to detect seven different objects which include pedestrian/ person, vehicles (car, bus, bike, and bicycle), animal (dog), and light/ sign poles. All these objects are commonly found on the roadside thus it will provide a better perspective for the driver's assistance. In this study, we have reviewed four large-scale datasets in the thermal domain. These datasets are available publicly and provide image sources with differing outdoor environmental and weather conditions. These datasets include the OSU Thermal pedestrian [16] database, KAIST Multi-Spectral dataset [32], FLIR ADAS dataset [36], and CVC-09 [34] datasets. Figure 1 shows the sample images from these datasets under different environmental and lighting conditions.

Table 1 provides the complete dataset attributes of all four datasets used in this study. The selected set of public datasets is used for optimal training and testing of four different network variants of YOLO-v5 named as X-large, large, medium, and small models. Most of these datasets are specifically gathered and proposed for autonomous driving applications. We have used data samples from four different public datasets as shown in table 1 for the training of all four network variants of YOLO-v5 architecture. The numeric performance comparison of all the model variants has been evaluated in the experimental results and discussion section thus summarizing the best models in terms of precise accuracy and lower inference time. The trained networks are tested on both public as well as locally gathered datasets.

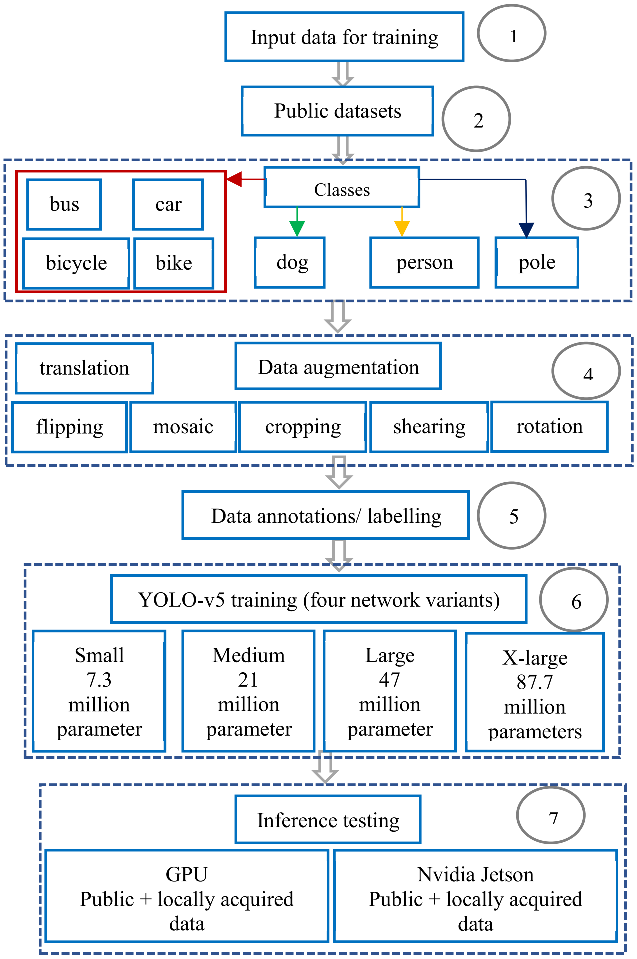

Input data for training
1
Public datasets
2
Classes
3
bus
car
bicycle
bike
dog
person
pole
translation
Data augmentation
4
flipping
mosaic
cropping
shearing
rotation
Data annotations/ labelling
5
YOLO-v5 training (four network variants)
6
Small 7.3 million parameter
Medium 21 million parameter
Large 47 million parameter
X-large 87.7 million parameters
Inference testing
7
GPU Public + locally acquired data
Nvidia Jetson Public + locally acquired data

FIGURE 2. Complete block diagram representation for object detection in thermal spectrum for driver assistance using end-to-end YOLO-v5 architecture.

A. TRAINING AND LEARNING APPROACH

In this work, we have included roadside objects for driver assistance comprising of seven different classes. Figure 2 shows the complete block diagram representation of the proposed methodology used in this study. It includes four different vehicles (i.e., bicycles, bikes (motorbikes), buses, cars), dogs in animal class, pedestrians or people, and roadside poles as shown in figure 2. The data samples from these classes are demonstrated in figure 3.

The individual class-wise training data distribution is shown in figure 4. A total of 32,715 data samples has been

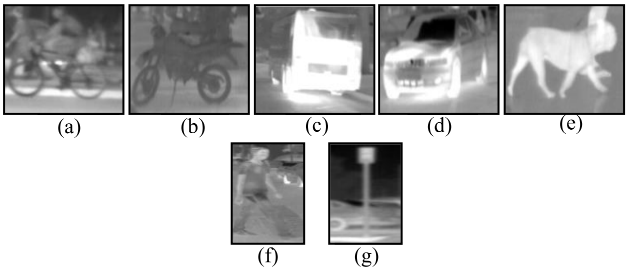


FIGURE 3. Seven different data classes for training YOLO-v5 in thermal spectrum a) bicycle, b) bike, c) bus, d) car, e) dog, f) person, g) sign/street pole.

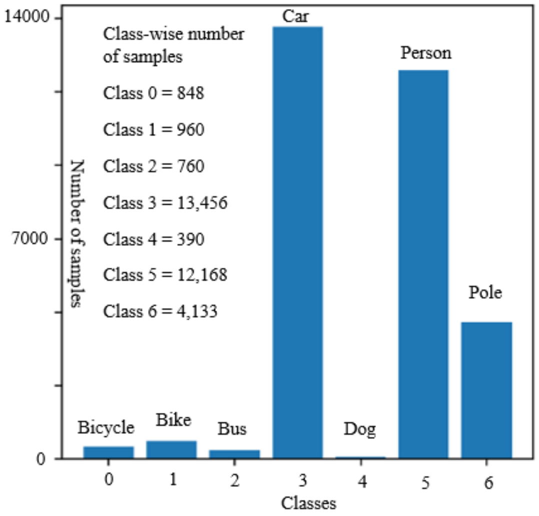


FIGURE 4. Class-wise distribution of complete thermal data used during the training process.

used along with their respective class labels in the training process of the YOLO-v5 framework. Figure 5 shows the data distribution of all the classes.

In this work, we have focused on using Ultralytics [8], [9] resource for the training of YOLO-v5 architecture on various public thermal datasets. During the training process, the configuration file is updated accordingly to our requirements specified in the head layer to adapt the number of classes (7 classes) on our dataset. To achieve precise training accuracy, we have trained all the network variants of YOLO-v5 architecture using both Stochastic Gradient Decent (SGD) [38] with the momentum of 0.9 as well as Adaptive learning rate optimization (Adam) [39] optimizer. SGD is considered a state-of-the-art optimizer for training deep convolution neural networks. It works by performing parameter update for each training batch rather than updating the whole training batch at once. It performs faster on large training samples. Moreover, it is computationally less expensive and has the ability to converge much faster as compared to batch Gradient Decent (GD) optimizer [40]. Adaptive learning rate optimization (Adam) algorithm work by taking advantage of adaptive learning rates thus computing the individual learning rate for individual parameters. It has various key benefits in comparison to other state-of-art optimizers which include invariant to diagonal rescale of the gradients, it requires less amount of memory, it is computationally more resourceful and lastly, it is more appropriate for noisy gradients. Lastly in this work rather than relying on one fixed learning rate we have used a one-cycle learning policy to find the optimal learning rate for our custom thermal training set. The main reason for using this method is to achieve robust results during the training process of complex network variants of YOLO-v5 architecture. The algorithm works by following the Cyclical Learning Rate (CLR) to achieve faster training time with the regularization effect.

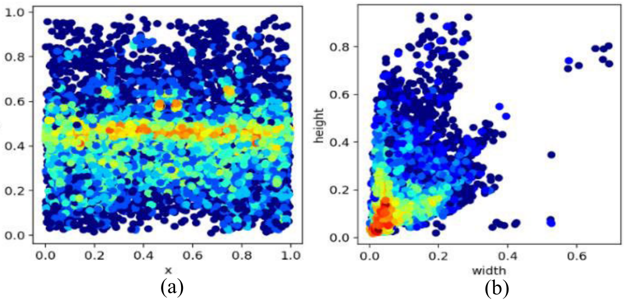


FIGURE 5. Training data distribution a) un-distributed data samples, b) class-wise clustered training data samples.

B. TRAINING DATA AUGMENTATION/TRANSFORMATION

Deep learning models are not considered ideal solutions with limited data options. To overcome this challenge, large data sets are required to perform optimal training of networks. Data augmentation is an effective way of producing many new training samples with diversity using the existing datasets. The synthetically generated data samples can be used with the original data to build large training sets. In this work, we have used various data transformation methods as shown in block 4 of figure 2. The further details of each augmentation method are as follows.

- Flipping: This method is used to perform image flipping in different directions which include up, down, left, and right directions. It helps the model to be insensitive to subject orientation.
- Rotation: This method is used to add variability to rotations thus helping the model to be more resilient to the camera roll.
- Image cropping: This technique is used to augment changeability to positioning and size. It helps the model to be more resilient to subject translations and camera positions.
- Image shearing: It is used to add shifting to the image by providing desired vertical and horizontal angles.
- Translation: It is used to move the image along the horizontal and vertical axes. This method of transformation is very useful as objects can be located almost anywhere in the image.
- Mosaic transformation: It is considered an advanced form of image augmentation operation. It works by

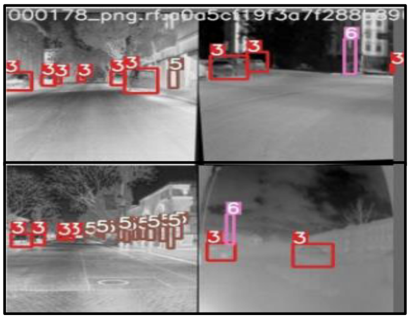


FIGURE 6. Mosaic transformation formed by combining four different training samples with different class labels including cars labeled as 3, person labeled as 5 and poles labeled as 6.

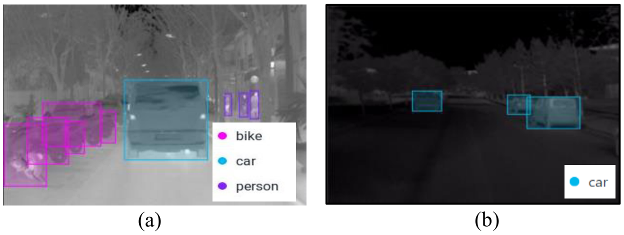


FIGURE 7. Annotated sample images with tight bounding boxes from two different datasets a) CVC-09 dataset sample (road-side view), b) KAIST dataset sample (road-side parking).

combining different training samples in one image with varying ratios. We have employed this augmentation method during the training process. It helps the model to learn how to identify the objects at a smaller scale than normal. It also encourages the model to localize different types of images in different portions of the frame. Figure 6 shows the mosaic transformation of four different training samples with different class labels.

C. DATA ANNOTATIONS

In this study, we have performed manual bounding-based annotations for all the thermal classes. Tight bounding box-based annotations were performed on all the frames for the training of YOLO-v5 framework. All the network variants are trained to detect and classify bicycles, bikes, buses, cars, dogs, pedestrians, and roadside poles in stimulating environmental conditions which include sunny weather, cloudy weather, night-time with total darkness, daytime, and other challenging environmental conditions. Table 2 shows the respective distribution of all the annotated frames selected from four different public datasets in varying environmental conditions.

Fig. 7 shows some of annotated training data samples in different environmental conditions which are selected from two different datasets and depicting different objects.

D. LOCALLY RECORDED TESTING DATA USING LWIR THERMAL CAMERA

The trained network variants are tested on both public as well as newly gathered test data to validate the efficacy of YOLO-V5 framework. The new test data is acquired using a prototype thermal camera specifically designed for this project. It is based on an uncooled micro-bolometer thermal camera array that embeds a France [41] Long Wave Infrared (LWIR) sensor. Figure 8 shows the images of the prototype thermal camera used in this research project [10]. Whereas table 3 shows the technical specifications of the uncooled thermal camera. The data is collected in two different approaches. In, the first approach the data is gathered in a stationary manner by placing the camera at a fixed place. The camera is mounted on the tripod stand at a fixed height of nearly 35 inches such that the roadsides objects are covered in the video stream. The thermal video stream is recorded at 30 Frames Per Second (FPS). The data is recorded in different weather conditions.

TABLE 2. Existing thermal dataset annotations.

Total frames selected from different public datasets	Total Annotated frames	Class wise annotations
1. CVC-09: 1650 frames 2. FLIR Adas: 1700 3. KAIST Multispectral dataset: 1700 4. OSU Thermal Dataset: 200 Total frames = 5,250	5100 frames with bounding box annotations	1. Bicycles: 848 2. Bikes: 960 3. Buses: 760 4. Cars: 13,456 5. Dogs: 390 6. Person: 12,168 7. Poles: 4,133 Total annotations = 32,715

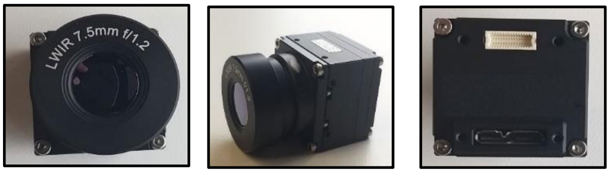


FIGURE 8. Uncooled LWIR prototype thermal camera images from different angles developed under the Heliaus EU project [37].

TABLE 3. Technical specifications of protoype LWIR thermal camera.

Prototype thermal camera specifications	
Quality and Type	VGA, Long Wave Infrared (LWIR)
Resolution	640 x 480 pixels
Focal length (f)	7.5 mm
F-Number	1.2
Pixel Pitch	17 µm
HFOV	90-degree, 890 mm

The data acquisition setup is shown in figure 9 whereas figure 10 shows the complete roadside view captured from logitech RGB Camera.

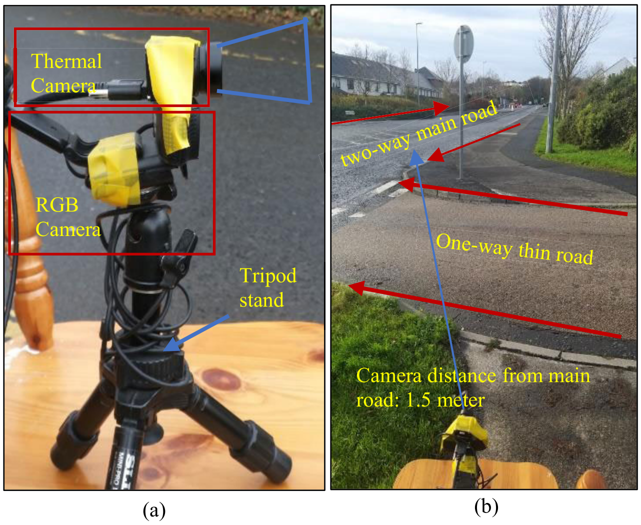


FIGURE 9. Data acquisition setup by mounting the camera at a fixed place and height, a) thermal and RGB camera mounted at a height of 0.9 meters from the ground surface, b) roadside field of view showing two-way main road and the one-way thin road joining the two-way main road.

FIGURE 10. Visual frame in the evening time with cloudy weather showing the complete roadside view.

In the second approach rather than collecting data in a static approach the camera is mounted in the car and data is collected through the car. The main reason for collecting data in two different approaches is to check the effectiveness of trained network variants of YOLO-v5 framework on diversified and distinctive local data in different weather and environmental conditions. Figure 11 shows the thermal camera along with the visible camera setup on the car.

Figure 12 shows the recorded sample thermal frames in the day, evening, and nighttime with different weather conditions using both by placing the camera at a fixed place and by mounting the camera on the car.

IV. EXPERIMENTAL RESULTS ON GPU AND EDGE DEVICE

This section will exhibit the thermal object detection results along with the performance comparison of four different network variants of YOLO-v5 framework. In this study, different testing approaches have been employed thus making a fair numeric comparison between these approaches on the test dataset. These methods along with their experimental results are further discussed in subsections. Moreover, in this study, we have deployed the smallest model variant (in terms of having the least number of model parameters), and yet the fastest network variant (in terms of least inference time) of YOLO-v5 framework on Nvidia Jetson edge device [11]. This will eventually help us in the form of trained model portability for diversified ADAS applications.

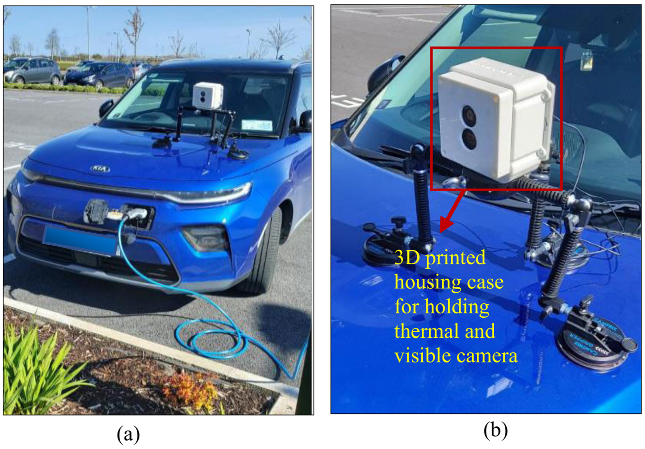

3D printed housing case for holding thermal and visible camera
(a)
(b)

FIGURE 11. Data acquisition setup by mounting the cameras on the car, a) thermal and RGB camera sealed in the white box and fixed on a suction tripod mount, b) closer view of the 3d printed white box holding thermal and visible cameras.

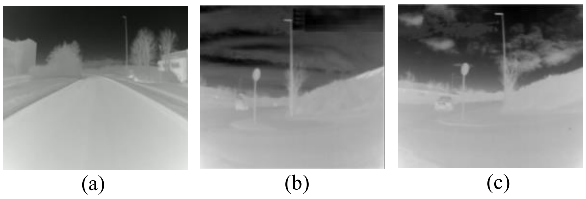


FIGURE 12. Recorded thermal data samples in different weather conditions by placing the camera on the car and by placing the camera at a fixed place, a) daytime with sunny weather, b) evening time with cloudy weather, c) nighttime with partially cloudy and windy weather.

In the initial phase of experimental results, we have used the pre-trained weights and tested them on thermal datasets without undergoing any training process. However, the results were not satisfactory as detector was unable to detect most of the objects in thermal frames. In the next phase, we have trained all the networks of YOLO-v5 from scratch and used newly trained model weights for evaluation on five different thermal test datasets (4 are public and 1 is local).

A. TRAINING CONFIGURATION AND TESTING APPROACHES

This section will mainly focus on the training configuration using two different optimizers which include SGD and Adam employed in this study and different testing approaches. The complete training configuration is provided in table 4. The training process is performed on a server-grade machine equipped with XEON E5-1650 v4 3.60 GHz processor, 32 GB of ram, and GEFORCE RTX 2080 graphical processing unit. It comes with 12 GB of dedicated graphical memory, memory bandwidth of 616 GB/second, and 4352 cuda cores. During the training process, the training batch size is fixed

TABLE 4. YOLO-v5 training configuration.

Hyperparameter Selection		
1	Initial Learning Rate	0.001
2	Final learning Rate	0.2
3	Warmup initial bias learning rate	0.1
4	Learning policy	One cycle learning rate
5	Training and Testing Batch size	8 & 32
6	Activation funtion	Sigmoid
7	Optimizer	SGD and Adam
8	Loss function	Binary Cross-Entropy (BCE) with Logits Loss
9	Epochs	100
10	Warmup epochs	3
11	Momentum	0.9
12	Warmup momentum	0.8
13	Weight decay	0.0005
14	IoU threshold	0.5
15	Anchor multiple threshold	4.0

to 4. The training process is performed on Pytorch deep learning framework [42]. It is important to mention that we have trained all the networks from scratch by un-freezing all the network layers and building new weights rather than transfer learning the networks to adapt the models for thermal data.

In the proposed study three different test-time approaches are employed which include test-time with no augmentation (TTNA), test-time augmentation (TTA), and test-time with model ensembling (TTME) methodology. The details of these are methods are as follows.

- TTNA: It is referred to as a conventional testing approach used for the unseen testing data provided to the trained object detection models. In this method, we don't perform any data augmentation/ transformation. Since no additional data augmentation operations are performed, the inference time depends on how dense the trained model is.
- TTA: Test-time augmentation is often helpful to achieve more robust results in the form of high inference accuracy from trained networks. Test-time augmentation is an extensive application of data augmentation applied to the test dataset. Specifically, it works by creating multiple augmented copies of each image in the test set, having the model make a prediction for each, then returning an ensemble of those predictions. However, since the test dataset is enlarged with artificially augmented images the inference time also increases as compared to TTNA which is one of the drawbacks of this approach. In this study, test-time augmentation method is performed on the test dataset by incorporating three different augmentation methods which include image shifting, cropping, and flipping.
- TTME: This is a technique for establishing the performance of multi-modal trained network variants on the test datasets. In machine learning model ensembling or ensemble learning refers to as using multiple trained networks at the same time in a parallel manner to produce one optimal predictive inference model. In this study, we have tested the performance of individually trained variants of the YOLO-v5 framework and selected the best combination of models which in turn helps in achieving better mean-average precision (mAP) scores on the validation set. However, as the trade-off, the individual inference time on each test frame/ image increases relatively as compared to TTNA methods.

B. TRAINING RESULTS

This section will summarize the training results of all the network variants of YOLO-v5 framework. The training accuracy and loss results are analyzed using different quantitative metrics to fully evaluate the effectiveness of all the trained models. The overall loss in the YOLO-v5 framework is calculated as compound lost based on three different scores which include objectness score, class probability score, and bounding box regression score. In this study, we have used Binary Cross-Entropy (BCE) with Logits Loss function in pytorch for loss calculation of class probability and object score. Whereas, the model accuracy is computed in terms of recall rate, model precision, and mean average precision (mAP). These accuracy metrics are explained below respectively.

1) RECALL AND PRECISION

In machine learning recall or sensitivity is counted as a critical statistical tool which is also referred to as true positive rate. It is defined as the ratio of true positive and the total amount of ground truth positives. The precision of any class is defined as the ratio of true positive (TP) and the sum of predicted positives. It is also referred to as positive predicted values. Equation (1) shows the formula of recall and precision metrics.

$$Recall = \frac{tp}{tp + fn} \times 100 \quad Precision = \frac{tp}{tp + fp} \times 100 \tag{1}$$

where tp is the true positives, fn is defined as false negatives and fp is the false positives.

2) MEAN AVERAGE PRECISION (MAP)

The mean average precision (mAP) is a standard metric used to measure the performance of deep learning models trained for applications such as information retrieval and object detection tasks. It is defined as the area under the Precision-Recall curve. The mAP for the object detection model is the average of the AP computed for all the classes. Equation (2) shows the formula for calculating the AP.

$$AP = \sum\nolimits_{Recall_i} Precision\,(Recall_i) = 1 \tag{2}$$

As mentioned earlier in (Section III-A), we have used both SGD and ADAM optimizers during the training process and selected the trained models with the best performance for validation on the public as well as locally gathered test data. Table 5 shows the area under the Precision-Recall curve for

TABLE 5. YOLO-v5 training results of all the four network variants.

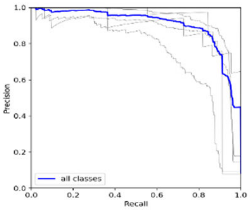
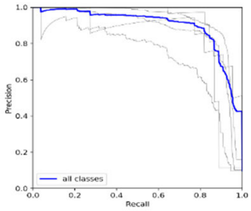
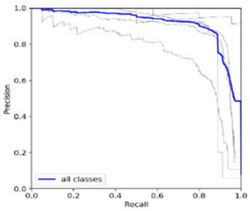
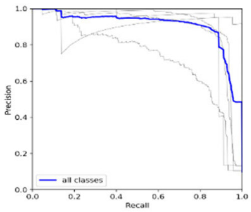
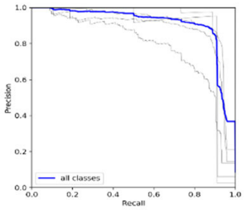
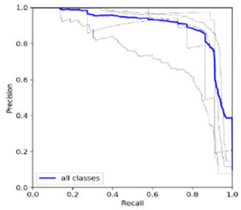
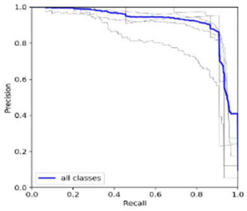
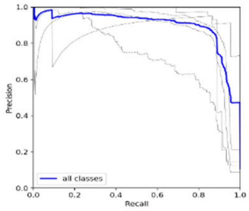
Precision-Recall Curve of all four network variants of YOLO-v5		
Model	Optimizer: SGD	Optimizer: Adam
Small	mAP = 90.43% 	mAP = 89.80% 
Medium	mAP = 91.26% 	mAP = 91.05% 
Large	mAP = 90.53% 	mAP = 89.57% 
X-large	mAP = 91.31% 	mAP = 90.30% 

TABLE 6. YOLO-v5 trained models accuracy and loss comparisons.

Optimizer	GPU usage/epoch (GB)	Training time required (Hours)	P %	R %	Losses		
					Bounding box loss	Objectness loss	Classification loss
Model: Small							
SGD	2.53	0.8	69.63	93.83	0.026	0.025	0.00079
ADAM	2.58	0.708	61.92	94	0.028	0.027	0.00095
Medium							
SGD	3.11	1.2	70.84	94.51	0.023	0.022	0.00070
ADAM	3.34	1.24	67.68	94.23	0.027	0.025	0.00115
Large							
SGD	6.65	1.8	75	93	0.021	0.020	0.00058
ADAM	6.68	1.83	68.92	92.51	0.025	0.025	0.00112
X-large							
SGD	9.74	3.2	75	92.89	0.020	0.019	0.00062
ADAM	9.82	3.25	69	93.76	0.024	0.023	0.00117

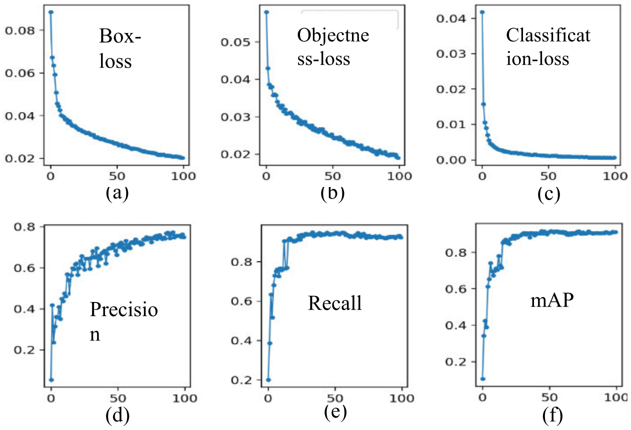


FIGURE 13. Training and Loss graphs of X-large model using SGD optimizer a) bounding-box loss, b) objectness loss, c) classification loss, d) model Precision, e) model recall curve, and f) mean average precision (mAP).

all the classes of four different network variants trained from scratch using both SGD and ADAM optimizer. Also, during the training process, we have made a comparative analysis of total graphical memory usage and the total training time required for all the models.

For a better understanding of the performance comparison of all the models, table 6 shows the numerical results of all the accuracy metrics, loss metrics, graphical memory usage, and overall training time required using both SGD and ADAM optimizer. It can be observed and summarized from table 5 and table 6 that models trained using SGD optimizer have performed significantly better as compared to models trained using ADAM optimizer in the terms of better mAP value, better precision scores, lower GPU usage, lower training time, and finally lower losses. Also, by analysing the individual performance of all the models, the X-large model has achieved the best mAP score of 91.31% with the lowest losses as compared to all other models. Whereas the medium network variant has scored the second-best mAP score of 91.26% along with the highest recall rate of 94.51% among all the models. In terms of the highest precision scores, the large and x-large model has outperformed all other models thus achieving the best precision score of 75%.

Figure 13 shows the accuracy and loss graphs of the x-large model as it has achieved exceptional performance in terms of the highest map scores and lower loss values when analysing the performance of other network variants of YOLO-v5 framework. However, as the trade-off, this model requires the highest training time and greater GPU usage

TABLE 7. Test data collected from public datasets.

Public Datasets	Weather Conditions	Environment	Frames selected	Total frames
1 CVC-09	Daytime and nighttime	Roadside	1000 + 1000	2436
2 FLIR-ADAS	Daytime and nighttime with sunny and cloudy weather conditions	Roadside	250	
3 KAIST Multispectral	Nighttime	Downtown	136	
4 OSU Thermal Dataset	Daytime with cloudy weather conditions	University campus	50	

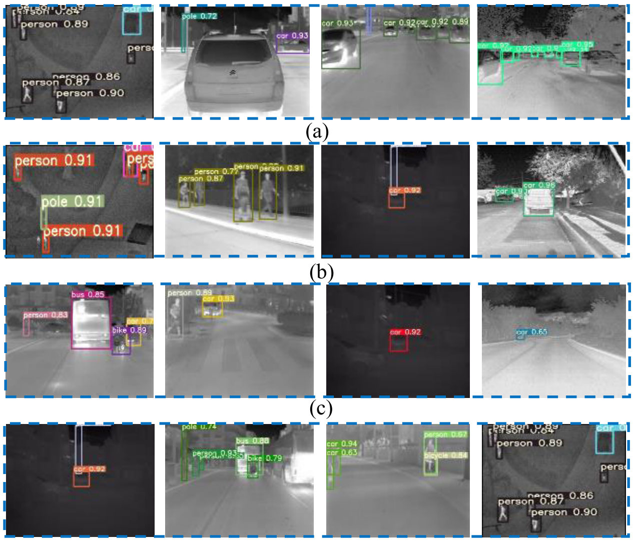
which makes it computationally more expensive. The overall model is consisting of 407 layers, 88.47 million gradients, and the final trained model weight size is 173 mb.

C. VALIDATION RESULTS ON PUBLIC AND LOCALLY GATHERED TEST-DATA

In the first phase, the performance of YOLO-v5 trained networks is evaluated on the unseen data gathered from publicly available datasets. As discussed earlier in (Section IV-A) three different test approaches are used which include TTNA, TTA, and TTME to validate the efficacy of four different networks of YOLO-v5 architecture. The training results signify that the SGD optimizer has performed better than the ADAM optimizer, however, during the testing phase, we have included the results extracted from trained networks using ADAM optimizer thus making fair testing evaluation among all the set of trained models. The test data comprises different weather conditions, different environments/ places, and varying distances of the objects from the camera. Table 7 shows the total number of frames used as the test data from four different publicly available datasets along with their respective attributes.

In the first segment, the inference test is run on test data using the TTNA approach. Whereas, in the second part we have run the inference results using the TTA approach. During the complete testing phase, the confidence threshold is set to 0.5. Figure 14 shows the sample results on sixteen different frames selected arbitrarily from the test-set using the TTNA approach and models trained using the SGD optimizer [38]. These frames consist of either single or multiple objects in the thermal spectrum from seven different classes as shown in figure 3. The test results are sub-divided into four parts extracted from four different network variants of the YOLO- v5 framework.

It can be observed from figure 14 that inference results on test data using different network variants trained on thermal data have improved significantly as compared to results when just using the pre-trained weights. However, by closely analysing the results still, we are unable to detect and classify some of the objects in thermal frames with challenges like scale and view-point variations, occlusions, and overlapping classes. For instance, in figure 14 (a) frame 2 smaller version of the thermal object detector is unable to detect the car close to the camera mounted on another car for recording the data. Similarly, in figure14 (b) frame 1 rather than detecting two cars, the medium model can detect only one car.



(a)

(b)

(c)

(d)

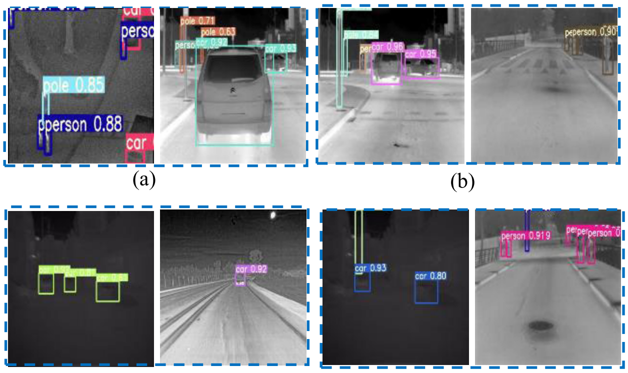
FIGURE 14. Object detection inference results with class confidence scores on sixteen different frames from four different public datasets using test-time without augmentation approach (TTNA) a) results extracted using small network variant, b) results extracted using medium network variant, c) results extracted using large network variant and d) results extracted using x-large network variant.

To overcome these issues and further improve the test accuracy, the inference test is run using the test-time augmentation (TTA) approach. The average inference time per frame using the TTNA method varies depending on the size of the model. The average inference time using the small model is 11 milliseconds whereas the average inference time using the x-large model is 21 milliseconds. Figure 15 depicts the inference results on eight different samples from the test data using TTA approach and four different network variants of the YOLO-v5 framework using SGD optimizer [38]. It can be observed that results are improved marginally as compared to the TTNA approach and significantly as compared to using originally pre-trained weight. This technique helps in detecting and classifying the object in the thermal spectrum more robustly with data complications such as occlusion, overlapping classes, scale variation, and varying environmental conditions. However, as the trade-off the average inference time per frame by employing TTA method increases as compared to TTNA method since additional augmentation operations are performed during the testing phase.

Table 8 shows the numerical performance comparison between TTNA and TTA for all the network variants by computing the mAP, precision, recall rate, and average inference time required per image. For this purpose, we have

TABLE 8. YOLO performance evaluation on test images from public datasets (The best value per metric is highlighted in green and emboldened).

Optimizer	Method Test time with no Augmentation (TTNA)				Method Test time Augmentation (TTA)			
	Precision	Recall %	mAP %	Average inference time/ image (millisecond)	Precision	Recall %	mAP %	Average inference time/ image (millisecond)
Model: Small								
SGD	88.4	82.5	79.5	11	83.5	89.8	85.4	23
ADAM	85.9	81.5	77.8	11	72.9	85.8	81.8	21
Model: Medium								
SGD	**90.5**	86.5	83.2	13	86.2	**90.1**	85.8	28
ADAM	86.5	81.4	78.3	13	82	87.5	83.7	28
Model: Large								
SGD	89.3	**87.2**	**84.1**	17	**86.5**	**90.2**	**86.6**	35
ADAM	84.7	81,6	78.1	18	80.5	87.8	83.3	37
Model: X-Large								
SGD	88.8	85.6	82	21	85.7	89.3	85.2	53
ADAM	86.2	76	72	23	80.2	86.7	81.5	54



(a) (b) (c) (d)

FIGURE 15. Object detection inference results with class confidence scores on eight different frames from four different public datasets using test-time with augmentation approach (TTA) a) results extracted using small network variant, b) results extracted using medium network variant, c) results extracted using large network variant and d) results extracted using x-large network variant.

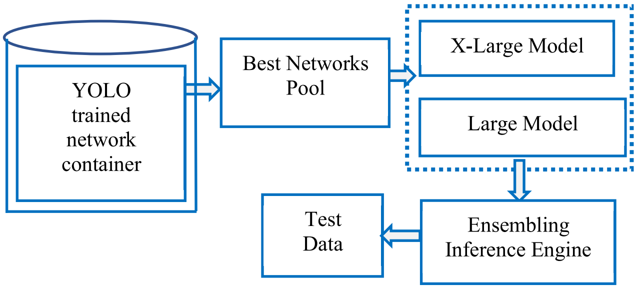


FIGURE 16. Model ensembling inference engine architecture.

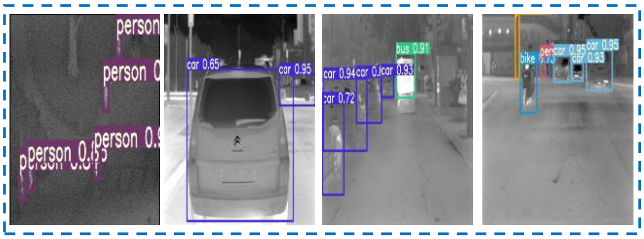


FIGURE 17. Model ensembling inference results on four different frames with multiple objects, overlapping classes, varying distance of the object from the camera, and different environmental and weather conditions.

short-listed a test-set of 250 frames with image complexities like occlusions, scale and viewpoint variations, and especially multiple objects with closely overlapping classes from the overall test data as shown. By analyzing the results from table 8, we can summarize that the large model using TTA method and SGD optimizer has achieved the best mean average precision, recall rate, and precision score of 86.6%, 90.2%, and 86.5% respectively as compared to other network variants of YOLO-v5 framework.

To further enhance the testing accuracy and explicitly reduce the inference time as compared to the TTA method a third testing approach i.e., model ensembling is used in this study. In this approach, we have tried various combinations of models by running them in ensembling style and selected a new set of two best models by evaluating their performance on test data as shown in figure 16.

As demonstrated in figure 16 the newly proposed ensembled inference engine consists of large and x-large models to produce one optimal predictive model. It is then evaluated on the test data as shown in table 7. Figure 17 shows the individual inference results on four selected frames with complex scenarios like multiple objects with overlapping classes, occlusions, object scale, and viewpoint variations, and different weather conditions.

Whereas, table 9 shows the numerical performance in terms of mAP, precision, recall rate, and average inference

TABLE 9. Model ensembling performance evaluation on test images from public datasets.

Optimizer	Method Test time with Model Ensembling (TTME)			
	Precision %	Recall %	mAP %	Average inference time required/ image (millisecond)
SGD	87.6	88.8	85.5	33

TABLE 10. Locally gathered test data from uncooled LWIR camera.

Method	Weather Conditions	Environment	Frames selected	Total frames
By mounting the camera at a fixed place	Daytime, evening time, and nighttime with cloudy and windy weather	Roadside	5000	20,000
By mounting the camera on the car	Daytime and evening time with sunny and cloudy weather	Cityside and University Campus	15000	

time required per image using the model ensembling method. In the second phase, we have used the same model ensembling approach to validate its robustness on locally gathered novel test data. As mentioned earlier in (Section III-D), test data is collected in two different methods which include mounting the camera at a fixed place, and in the second method, data is gathered by mounting the camera on the electric car and driving it around the university campus and city side. The locally gathered test data comprises different weather conditions, different environments/ places, different lighting conditions (day, evening and night time), multiple objects of different classes, and scale variations.

Table 10 shows the total number of short-listed frames used as the test data from our locally generated dataset along with their respective attributes.

Figure 18 displays the inference results on eight distinct thermal frames gathered from the uncooled prototype LWIR camera used in this project. It can be observed that an proposed ensemble inference engine comprising of x-large and large network variant has achieved precise results as it can detect and classify multiple objects of different classes in newly gathered local test data.

However, to further investigate its effectiveness on local test data, we have computed various accuracy metrics which include precision, recall, mAP, and mean inference time required per image as it was computed in the case of public datasets on a set of 250 thermal frames. These results are shown in table 11.

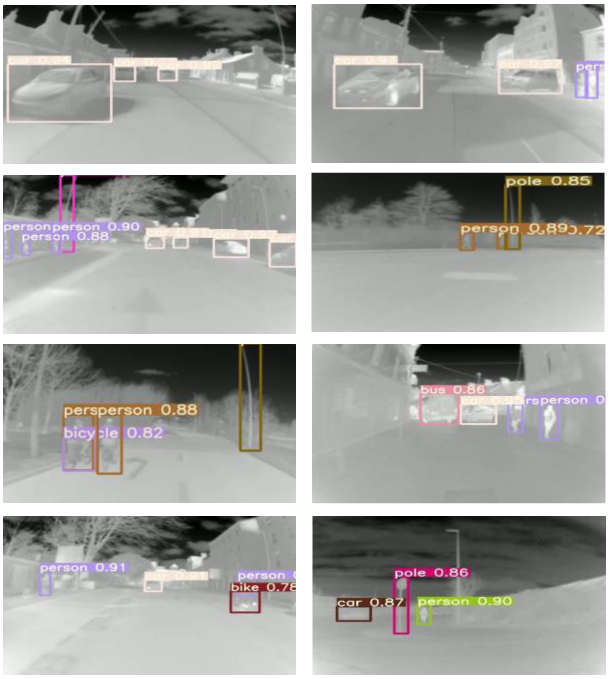


FIGURE 18. Model ensembling inference results on eight different frames acquired from prototype thermal camera with multiple objects, overlapping classes, the varying distance of the object from the camera, and different environmental and weather conditions.

TABLE 11. Model ensembling performance evaluation on test images from the locally gathered dataset.

Optimizer	Method Test time with Model Ensembling (TTME)			
	Precision %	Recall %	mAP %	Average inference time required/ frame (millisecond)
SGD	83.5	78.1	70	29

D. DEPLOYMENT AND VALIDATION RESULTS ON EDGE COMPUTING

After successful convergence and testing of YOLO networks on GPU architecture, in the next step, we drive towards the deployment of the trained network on edge-inference architecture. The primary goal is to create a flexible, scalable, secure, and more automated hardware system thus allowing us to easily export the trained network weights for easy model portability. It will be beneficial when integrating the edge devices with thermal camera for deploying it in intelligent automotive sensor suite for real time analysis. The core benefits of deploying the trained machine learning (ML) model on edge devices include.

1. The edge hardware is assumed to be more energy-efficient since it requires less amount of power resources as compared to single or clustered-based CPU and GPU server machines.

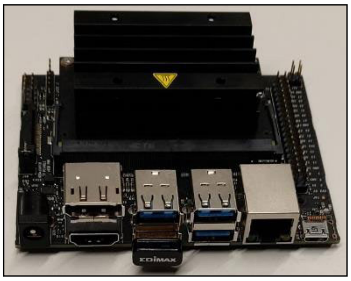


FIGURE 19. Nvidia Jetson nano developer kit for deploying the YOLO-v5 trained networks.

2. Locating and processing inference at the edge architecture lies in saving communication power.
3. Finally, the cost of edge-based inference hardware is considerably less as compared to other computational hardware such as field-programmable gate arrays (FPGA) and GPUs.

For this study, Nvidia Jetson Nano [11] developer kit is selected to evaluate the performance of YOLO-v5 trained model on thermal test data. It is a small yet powerful edge computer that allows us to run multiple neural networks in parallel manner for several computer vision applications such as image classification, object detection, segmentation, and speech processing. It is considered as all in an easy-to-use platform that runs in as little as five watts of power. Figure 19 shows the Nvidia Jetson nano developer kit equipped with CPU QUAD-core ARM A57 at 1.43 GHz and GPU 128-core Maxwell. It comes with a memory of 4 GB, 64-bit, LPDDR4 25.6 GB/s. Jetson nano has a total of four USB 3.0 ports, HDMI port, an ethernet port, and a barrel connector to power it via five volts and 4-ampere supply.

The small network variant of the YOLO-v5 framework is selected for further optimization as it has the minimum number of model parameters and requires the least inference time as compared to other models of YOLO which makes it computationally less expensive and cost-effective model. For better optimization, and further reduce the inference time we have used TensorRT inference optimizer [43]. It is a type of deep learning inference optimizer and runtime engine that delivers low latency and high throughput for deep learning inference applications. TensorRT-based inference engines can perform up to 40× faster than CPU-only platforms. The optimized inference models can be easily deployed to hyper-scale data centres, embedded and edge devices, and automotive product platforms. Figure 20 shows the adapted structural architecture design for converting the YOLO-v5 small deep learning model converged on thermal data to an optimized inference engine. As shown in figure 20 the process overflow for converting the trained network variant of the YOLO-v5 framework to TensorRT based optimized inference engine is consists of six main steps. In the first step, it maximizes throughput by quantizing models to 8-bit integer data type while preserving the accuracy. In the second step, it improves the use of GPU memory and bandwidth by fusing nodes in a kernel. In the next step, it performs Kernal auto-tuning. In the fourth step, it minimizes memory footprints and re-uses memory for tensors efficiently. In the last steps, it processes multiple input streams in parallel and finally optimizes neural networks periodically with dynamically generated kernels [43]. Once the model is optimized successfully, it is serialized and deserialized to run the inference test.

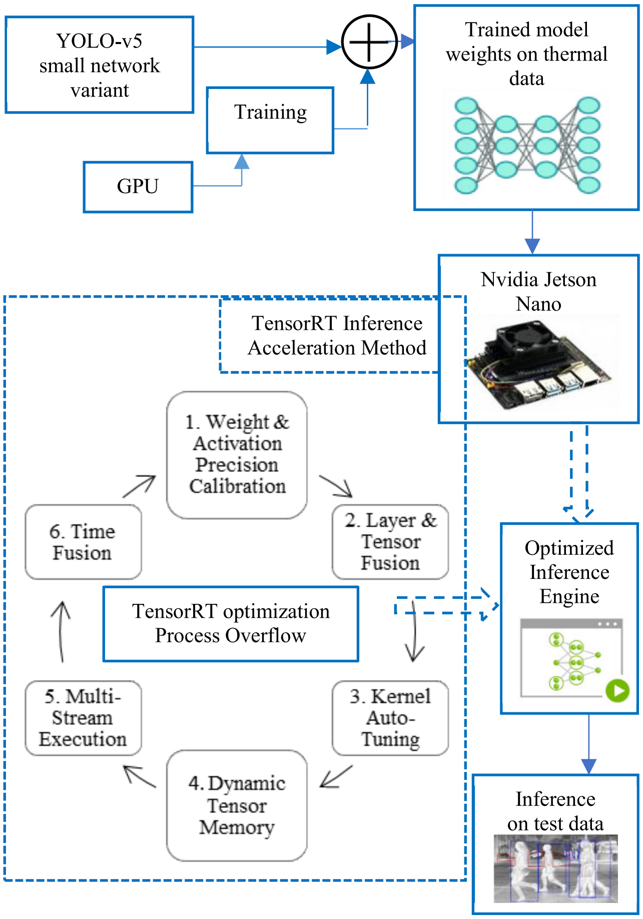
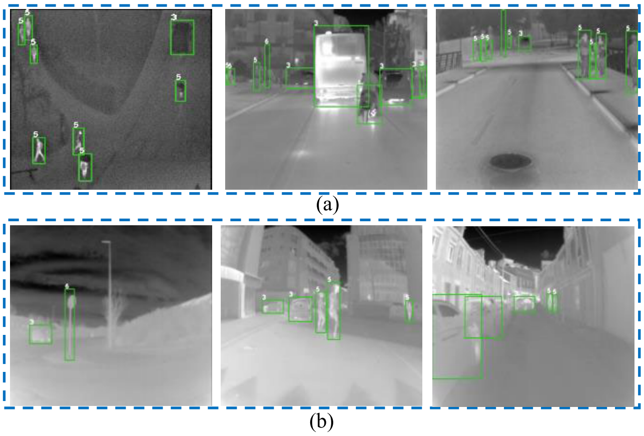

YOLO-v5 small network variant
Training
GPU
Trained model weights on thermal data
Nvidia Jetson Nano
TensorRT Inference Acceleration Method
1. Weight & Activation Precision Calibration
2. Layer & Tensor Fusion
3. Kernel Auto-Tuning
4. Dynamic Tensor Memory
5. Multi-Stream Execution
6. Time Fusion
TensorRT optimization Process Overflow
Optimized Inference Engine
Inference on test data

FIGURE 20. Structural architecture for converting YOLO-v5 trained network to TensorRT optimized inference engine.

Fig. 21 shows the inference results on six different thermal frames from the public as well as locally gathered test data. The generated results are in the form of bounding boxes with the respective class number. Figure 22 shows the test data results on 400 images in the form precision-recall curve for all the classes from both local and public test data on Jetson nano. Table 12 shows the comparative analysis of the inference time per frame and FPS rate between the typically trained model tested on GPU, optimized/ accelerated version of the model tested on GPU, and accelerated version of the model tested on Nvidia Jetson nano.

It can be observed from table 12 that inference time has reduced to nearly 55% on GPU by using the optimized version of the YOLO-v5 model through TensorRT which will eventually benefit us when running the inference test with a large number of test frames and subsequently running



(a)

(b)

FIGURE 21. Inference results on Nvidia Jetson nano using TensorRT optimization engine by showing respective class numbers, a) inference results on public datasets, b) inference results on the locally gathered test dataset.

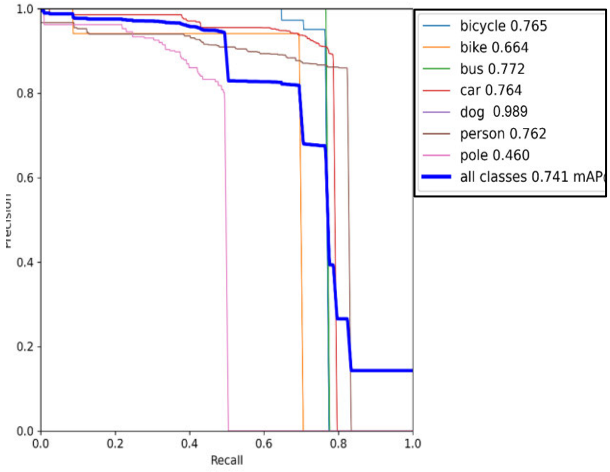


FIGURE 22. Test data results for all the classes on Jetson Nano using small network variant with an overall mAP of 74.1%.

TABLE 12. Inference time and FPS comparison on GPU and Jetson Nano.

Model: YOLO-v5 Small Variant		
Average inference time (milliseconds) /image and frames per second (FPS) comparison chart on GPU and Jetson Nano		
Non-optimized version on GPU	Optimized version on GPU	Optimized version on Nvidia Jetson Nano
11 ms	5 ms	320 ms
≈ 85 FPS	≈ 170 FPS	≈ 3 FPS

the inferences on higher frames per second (FPS) videos. Whereas on Nvidia Jetson it requires nearly 320 milliseconds average inference time per frame and FPS rate of 3 with an image resolution of 640 × 640 pixels.

V. DISCUSSION/ANALYSIS

This section will mainly emphasize individual training and testing performance comparison of all the model variants of YOLO framework.

- The small variant requires the lowest inference time among all other models with the mAP score of 85.4% using TTA approach which is nearly equal to the mAP score of the model ensembling method that is 85.5% during the testing phase.
- The medium model tends to achieve the best precision score of 90.5% using TTNA method and the best recall score of 90.1% using the TTA method among all other models during the testing phase. However, this model was unable to achieve robust mAP scores during both the training and testing phases.
- The large variant proves to the best network by achieving the highest mAP scores using both TTNA and TTA methods of 84.1% and 86.6% respectively during the testing phase as compared to other trained networks. However, it requires a longer inference time specifically when using the models with TTA method which is nearly 35 milliseconds per frame.
- The X-large model turns out to be the best-trained model thus scoring the highest mAP score of 91.31% and lowest losses using the SGD optimizer but during the testing phase, the model was unable to achieve exceptional accuracy on the validation/ test set along with the highest inference time using both TTNA and TTA methods. However, this model comes up with the best possible match with large network variant in the model ensembling method.
- By examining the overall performance of all the models, we can conclude that test accuracy using the TTA approach is significantly better as compared to TTNA. However, as a trade-off, the TTA method requires huge inference time which makes it a computationally more expensive method.
- Figure 23 shows the maximum and minimum inference time comparison chart of all the trained models tested on GPU using three different testing approaches. It can be observed that small network variant requires the least inference time using both TTA and TTNA methods thus making this model computably the least expensive and ideal network for real-time deployments especially on edge devices with comparatively less computational power. Also, the minimum and maximum time required by TTME approach is smaller as compared to TTA method which makes it more time-efficient networks.
- TensorRT optimizer has been used to further speed up the deep learning inference using Thermal-YOLO small network variant on GPU as well as edge embedded platform Nvidia Jetson Nano [11]. However, the performance of the optimized model is much superior on GPU as compared to Jetson Nano which is evidenced by the fact that the ratio of FPS rate between Jetson nano and GPU is 3:170. Also, the average inference time per frame on Jetson nano is nearly 98% more as compared to inference time on GPU.
- Lastly, the newly trained model weights performed much better as compared to pre-trained weights

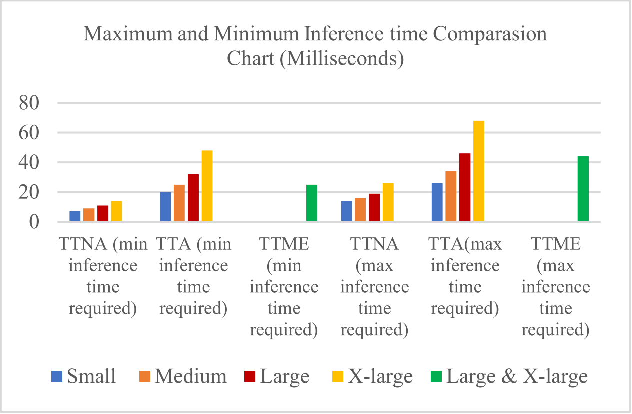
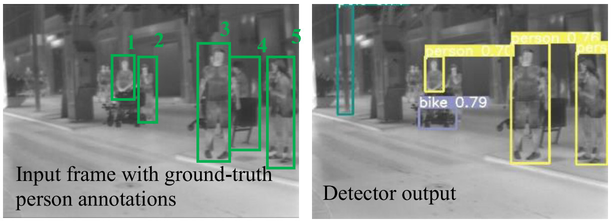


FIGURE 23. Inference time comparison chart of the trained models on GPU using test-time with no augmentation, test-time with augmentation, and test-time with model ensembling methods.

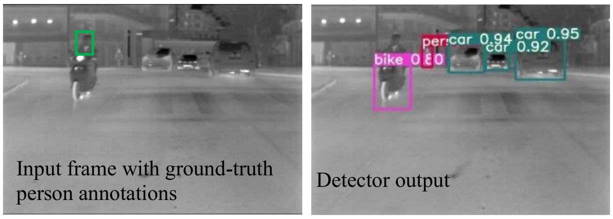
however still in some of the most complex thermal frames, models seem to provide inadequate results which are discussed and shown in some of the cases as follows.

A. CASE 1

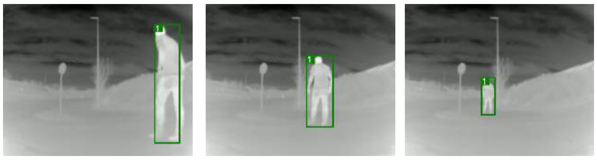


These results are obtained using the small network variant using TTNA method. It can be observed from the left side input frame that five people can be seen from a human perspective marked with manually annotated green boxes for better understanding. The right-side frame shows the detector output. The network was only able to detect three people since the second person was putting a hand in front of her face whereas the fourth person view was a side pose with occluded vision thus detector was unable to detect the second and fourth person. Moreover, we can see a second person holding a baby walker however the detector miss-classified it as a bike.

B. CASE 2



This result is obtained using the large variant using TTA method. There is a total of five objects in this frame (three cars and two people). The detector was able to detect four objects with good confidence scores however, it was failed in detecting one other person as demonstrated in the left side frame. This is because the person is riding the bike with occluded vision (wearing the helmet) which makes his facial features obstructed thus the model fails to detect and classify it.



(a) (b) (c)

FIGURE 24. Inference results along with person tracking using DeepSORT [44] by assigning id 1 a) frame 1, b) frame 15, c) frame 23.

VI. CONCLUSION/FUTURE WORK

In this work, we have proposed smart thermal perception systems effective for all lighting conditions using AI-based object detection pipeline for the automotive sensor suite. Four different network variants of the YOLO-v5 framework have been employed and trained using four different public datasets using SGD as well as ADAM optimizer. The X-Large model turns out to be the best-trained model thus achieving the highest mean average precision. The performance estimation of trained network variants is validated using both public as well as locally gathered new test data in different weather and environmental conditions. The Large network variant comprising 47.4 million parameters has achieved the best mAP score of 84.1% using TTNA and 86.6% using the TTA method. To further reduce the inference time as compared to the TTA method without compromising the accuracy, the test-time with model ensembling (TTME) methodology has been used. X-large and large model proves to be the best network coupler thus producing the results as one optimal inference engine. The proposed model ensembling-based thermal inference engine achieves the overall mAP of 85.5% on test data accumulated from public datasets and 70% on locally gathered test data respectively. Secondly, we have used TensorRT optimizer to further reduce the model inference time that can eventually help in real-time deployments, especially on edge devices. The optimized version of the smaller trained model is tested on both GPU as well as Jetson nano thus getting 170 FPS on GPU and 3 FPS on jetson nano.

As the possible future directions, these systems can be deployed on more powerful edge devices with a higher flop rate and less operating power for optimal performance, especially in real-time environments. Moreover, we intend to include more thermal classes thus making the overall system more mature and robust. In addition to this, we can further integrate the current object detection system with object tracking thus estimating the position of an object, as well as incorporating position predicted by dynamics.

This will eventually help us in counting the number of vehicles, pedestrians, etc., along with their approximate estimated distance. One such example is demonstrated in figure 24 where we have integrated deep association metrics [44] tracking with YOLO-v5 for person tracking on three different thermal frames selected from locally gathered test data.

ACKNOWLEDGMENT

The authors would like to acknowledge the experts from Xperi-Ireland and Christopher Dainty for providing the support in preparing the data accusation setup and helping throughout in data collection and Quentin Noir from Lynred France for giving their feedback. Moreover, they would like to acknowledge the contributors of all the public datasets for providing the image resources to carry out this research work and ultralytics for sharing the YOLO-v5 Pytorch version.

REFERENCES

[1] V. John, S. Mita, Z. Liu, and B. Qi, "Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks," in *Proc. 14th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2015, pp. 246–249, doi: 10.1109/MVA.2015.7153177.

[2] I. J. Xique, W. Buller, Z. B. Fard, E. Dennis, and B. Hart, "Evaluating complementary strengths and weaknesses of ADAS sensors," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–5, doi: 10.1109/VTCFall.2018.8690901.

[3] F. Munir, S. Azam, M. A. Rafique, A. M. Sheri, M. Jeon, and W. Pedrycz, "Exploring thermal images for object detection in underexposure regions for autonomous driving," 2020, *arXiv:2006.00821*.

[4] M. Kristo, M. Ivasic-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using Yolo," *IEEE Access*, vol. 8, pp. 125459–125476, 2020, doi: 10.1109/ACCESS.2020.3007481.

[5] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft COCO: Common Objects in Context* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8693. Cham, Switzerland: Springer, 2014, doi: 10.1007/978-3-319-10602-1_48.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[7] N. U. Huda, B. D. Hansen, R. Gade, and T. B. Moeslund, "The effect of a diverse dataset for transfer learning in thermal person detection," *Sensors*, vol. 20, no. 7, p. 1982, Apr. 2020, doi: 10.3390/s20071982.

[8] (Accessed: Oct. 22, 2021). *YOLO-V5 GitHub Repository*. [Online]. Available: https://github.com/ultralytics/yolov5

[9] (Accessed: Oct. 26, 2021). *YOLO-V5 Website Resource*. [Online]. Available: https://zenodo.org/badge/latestdoi/264818686

[10] (Accessed: Oct. 28, 2021). *Heliaus European Union Project*. [Online]. Available: https://www.heliaus.eu/

[11] (Accessed: Oct. 14, 2021). *Nvidia Jetson Nano*. [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano-developer-kit

[12] R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, and A. Weller, "Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1–9, doi: 10.24963/ijcai.2017/661.

[13] D. Olmeda, C. Premebida, U. Nunes, J. M. Armingol, and A. de la Escalera, "Pedestrian detection in far infrared images," *Integr. Comput.-Aided Eng.*, vol. 20, no. 4, pp. 347–360, Aug. 2013, doi: 10.3233/ICA-130441.

[14] B. Besbes, A. Rogozan, A.-M. Rus, A. Bensrhair, and A. Broggi, "Pedestrian detection in far-infrared daytime images using a hierarchical codebook of SURF," *Sensors*, vol. 15, no. 4, pp. 8570–8594, Apr. 2015, doi: 10.3390/s150408570.

[15] R. Lahmyed, M. El Ansari, and A. Ellahyani, "A new thermal infrared and visible spectrum images-based pedestrian detection system," *Multimedia Tools Appl.*, vol. 78, no. 12, 2019, doi: 10.1007/s11042-018-6974-5.

[16] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *Proc. IEEE Workshops Appl. Comput. Vis. (WACV)*, vol. 1, Jan. 2005, pp. 364–369.

[17] A. Torabi, G. Massé, and G.-A. Bilodeau, "An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications," *Comput. Vis. Image Understand.*, vol. 116, no. 2, pp. 210–221, Feb. 2012, doi: 10.1016/j.cviu.2011.10.006.

[18] T.-T. Dai and Y.-S. Dong, "Introduction of SVM related theory and its application research," in *Proc. 3rd Int. Conf. Adv. Electron. Mater., Comput. Softw. Eng. (AEMCSE)*, Apr. 2020, pp. 230–233, doi: 10.1109/AEMCSE50948.2020.00056.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, *SSD: Single Shot MultiBox Detector* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9905. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-46448-0_2.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[22] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020, doi: 10.1109/TPAMI.2018.2844175.

[24] J. Liu, S. Zhang, S. Wang, and D. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 73.1–73.13, doi: 10.5244/c.30.73.

[25] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *Proc. 24th Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, Belgium, Apr. 2016, pp. 509–514.

[26] M. Vandersteegen, K. Van Beeck, and T. Goedemé, *Real-Time Multispectral Pedestrian Detection with a Single-Pass Deep Neural Network* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10882. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-319-93000-8_47.

[27] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 243–250, doi: 10.1109/CVPRW.2017.36.

[28] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 988–997, doi: 10.1109/CVPRW.2019.00130.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[30] M. Ivašić-Kos, M. Krišto, and M. Pobar, "Human detection in thermal imaging using YOLO," in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*. 2019, pp. 20–24, doi: 10.1145/3323933.3324076.

[31] C. Herrmann, T. Müller, D. Willersinn, and J. Beyerer, "Real-time person detection in low-resolution thermal infrared imagery with MSER and CNNs," in *Electro-Optical and Infrared Systems: Technology and Applications*, vol. 9987. Bellingham, WA, USA: SPIE, 2016, doi: 10.1117/12.2240940.

[32] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "KAIST multi-spectral day/night data set for autonomous and assisted driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 934–948, Mar. 2018, doi: 10.1109/TITS.2018.2791533.

[33] Q. Liu, Z. He, X. Li, and Y. Zheng, "PTB-TIR: A thermal infrared pedestrian tracking benchmark," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 666–675, Mar. 2020, doi: 10.1109/TMM.2019.2932615.

[34] Y. Socarrás, S. Ramos, D. Vázquez, A. M. López, and T. Gevers, "Adapting pedestrian detection from synthetic to far infrared images," in *Proc. Conf. Workshop Vis. Domain Adaptation Dataset Bias (ICCV)*, vol. 3, Sydney, NSW, Australia, Dec. 2013.

[35] Z. Wu, N. Fuller, D. Theriault, and M. Betke, "A thermal infrared video benchmark for visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 201–208, doi: 10.1109/CVPRW.2014.39.

[36] (Accessed: Oct. 22, 2021). *FLIR Thermal Dataset*. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/

[37] (Accessed: Oct. 31, 2021). *Darknet: Open-Source Neural Networks in C*. [Online]. Available: https://pjreddie.com/darknet/

[38] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT*. 2010, doi: 10.1007/978-3-7908-2604-3_16.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.

[40] (Accessed: Oct. 20, 2021). *Difference Between Batch Decent and Gradient Decent*. [Online]. Available: https://www.geeksforgeeks.org/difference-between-batch-gradient-descent-and-stochastic-gradient-descent/

[41] (Accessed: Oct. 20, 2021). *Lynred France*. [Online]. Available: https://www.lynred.com

[42] (Accessed: Oct. 14, 2020). *Pytorch Deep Learning Framework*. [Online]. Available: https://pytorch.org/

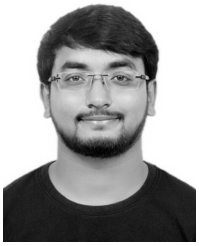
[43] (Accessed: Oct. 14, 2021). *Nvidia TensorRT for Developers*. [Online]. Available: https://developer.nvidia.com/tensorrt

[44] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649, doi: 10.1109/ICIP.2017.8296962.

MUHAMMAD ALI FAROOQ received the B.E. degree in electronic engineering from IQRA University, in 2012, and the M.S. degree in electrical control engineering from the National University of Sciences and Technology (NUST), in 2017. He is currently pursuing the Ph.D. degree with the National University of Ireland Galway (NUIG). He is also working with NUIG, one of the consortium partners in the Heliaus (thermal vision augmented awareness) Project funded by EU. His research interests include machine vision, computer vision, video analytics, and sensor fusion. He has won the prestigious H2020 European Union (EU) Scholarship.

PETER CORCORAN (Fellow, IEEE) currently holds the Personal Chair in electronic engineering with the College of Science and Engineering, National University of Ireland Galway (NUIG). He was the Co-Founder of several start-up companies, notably FotoNation (currently the Imaging Division, Xperi Corporation). He has more than 600 cited technical publications and patents, more than 120 peer-reviewed journal articles, and 160 international conference papers, and a co-inventor on more than 300 granted U.S. patents. He is an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He is also a member of the IEEE Consumer Technology Society for more than 25 years and the Founding Editor of *IEEE Consumer Electronics Magazine*.

COSMIN ROTARIU received the M.S. degree in medical informatics from the National University of Ireland Galway (NUIG), in 2006. He is currently working as a Senior Staff Engineer with the Systems Team, Xperi Corporation, Ireland. He is associated with industry for more than 12 years and holds vast experience in the areas of embedded systems designs, data filtering, compression, code optimizations, hardware design, and firmware designs. He is involved in many cutting-edge technology projects with industry which includes driver and occupant monitoring systems (DOMS), wireless networks, and low-level protocols designs for medical devices.



WASEEM SHARIFF received the B.E. degree in computer science from the Nagarjuna College of Engineering and Technology (NCET), in 2019, and the M.S. degree in computer science specializing in artificial intelligence from the National University of Ireland Galway (NUIG), in 2020. He is currently working as a Research Assistant with NUIG. He is associated with a research group in NUIG working on Heliaus (thermal vision augmented awareness) Project. His research interests include machine learning utilizing deep neural networks for computer vision applications, including working with visible, synthetic data, thermal data, and biosensors.

Appendix F

Evaluation of Thermal Imaging on Embedded GPU Platforms for Application in Vehicular Assistance Systems

Authors' Contribution to [19]

Contribution Criteria	*Contribution Percentage*
Research Hypothesis	MAF: 80%, WS: 20%
Experiments and Implementation	MAF: 100%
Background	MAF: 70%, WS: 30%
Manuscript Preparation	MAF: 70%, WS: 20% PC: 10%

Evaluation of Thermal Imaging on Embedded GPU Platforms for Application in Vehicular Assistance Systems

Muhammad Ali Farooq, Waseem Shariff, and Peter Corcoran, *Fellow, IEEE*

***Abstract*—This study is focused on evaluating the real-time performance of thermal object detection for smart and safe vehicular systems by deploying the trained networks on GPU & single-board EDGE-GPU computing platforms for onboard automotive sensor suite testing. A novel large-scale C3I Thermal Automotive dataset comprising of >35,000 distinct frames is acquired, processed, and open-sourced in challenging weather and environmental scenarios. The dataset is recorded from a lost-cost yet effective uncooled LWIR thermal camera, mounted stand-alone and on an electric vehicle to minimize mechanical vibrations. The state-of-the-art YOLO-v5 networks variants are trained using four different public datasets as well newly acquired local dataset for optimal generalization of DNN by employing SGD optimizer. The effectiveness of trained networks is validated on extensive test data using various quantitative metrics which include precision, recall curve, mean average precision, and frames per second. The smaller network variant of YOLO is further optimized using TensorRT inference accelerator to explicitly boost the frames per second rate. Optimized network engine increases the frames per second rate by 3.5 times when testing on low power edge devices thus achieving 11 fps on Nvidia Jetson Nano and 60 fps on Nvidia Xavier NX development boards.**

***Index Terms*—ADAS, object detection, thermal imaging, LWIR, CNN, edge computing.**

I. Introduction

THERMAL imaging is the digital interpretation of the infrared radiations emitted from the object. Thermal imaging cameras with microbolometer focal plane arrays (FPA) is a type of uncooled detector that provides low-cost solutions for acquiring thermal images in different weather and environmental conditions. These cameras when integrated with AI-based imaging pipelines can be used for various real-world applications. In this work, the core focus is to design an intelligent thermal object detection-based video analysis system for automotive sensor suite application that should be effective in all light conditions thus enabling safe and more reliable road journeys. Unlike other video solutions such as visible imaging which mainly relies on reflected light thus having the greater chances of being blocked by visual impediments, thermal imaging does not require any external lighting conditions to capture quality images and it can see through visual obscurants such as dust, light fog, smoke, or other such occlusions. Moreover, the integration of AI-based thermal imaging systems can provide us with a multitude of advantages from better analytics with fewer false alarms to increased coverage, provide redundancy and, higher return on investment.

In this research work, we have focused on utilizing thermal data for designing efficient AI-based object detection and classification pipeline for Advanced Driver-Assistance Systems (ADAS). Such type of thermal imaging-based forward sensing (F-sense) system is useful in providing enhanced safety and security features thus enabling the driver to better scrutinize the complete road-side environment. For this purpose, we have used state-of-the-art end-to-end deep learning framework YOLO-v5 on thermal data to predict 6 distinct objects, which include person, car, street-pole, bike, bicycle, and bus. In the first phase, a novel thermal dataset (https://github.com/Mali-Farooq/Thermal-YOLO) is acquired for training and validation purposes of different network variants of YOLO-v5. The data is captured using a prototype low-cost microbolometer based uncooled LWIR thermal camera with a resolution of 640x480, specifically designed under the ECSEL Heliaus research project [1]. The raw thermal data is processed using shutterless camera calibration, automatic gain control, bad-pixel removal, and temporal denoising methods.

Furthermore, the trained network variants are deployed and tested on two state-of-the-art embedded GPU platforms, which include NVIDIA Jetson nano [2] and Nvidia Jetson Xavier NX [3]. Thus, studying the extensive real-time and on-board feasibility in terms of various quantitative metrics, inference time, FPS, and hardware sensor temperatures.

The core contributions of the proposed research work are summarized below:

- Preparation and annotation of a large-scale C3I thermal automotive open-access dataset captured in different weather and environmental conditions.

Manuscript received October 25, 2021; revised February 8, 2022; accepted March 3, 2022. This work was supported by ECSEL Joint Undertaking (JU) under Grant Agreement 826131 (Heliaus project https://www.heliaus.eu/). The JU receives support from European Union's Horizon 2020 Research and Innovation Program and National funders from France, Germany, Ireland (Enterprise Ireland, International Research Fund), and Italy. *(Corresponding author: Muhammad Ali Farooq.)*

The authors are with the National University of Ireland Galway, College of Science and Engineering Galway, H91TK33 Galway, Ireland (e-mail: m.farooq3@nuigalway.ie; waseem.shariff@nuigalway.ie; peter.corcoran@nuigalway.ie).

Data is available on-line at https://github.com/Mali-Farooq/Thermal-YOLO

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2022.3158094.

Digital Object Identifier 10.1109/TIV.2022.3158094

TABLE I
EXISTING SOA THERMAL DATASETS

Datasets	Condition		Annotations	Objects	Total no. of frames	Image Resolution
	Day	Night				
OSU Thermal [4]	✓	✓	-	Person, Cars, Poles	284	360 X 240
CVC [5]	✓	✓	-	Person, Cars, Poles, Bicycle, Bus, Bikes	11K	640 X 480
LITIV [6]	-	-	-	Person	6K	320 X 240
TIV [7]	-	-	-	Person, Cars, Bicycle, Bat	63K	1024 X 1024
SCUT [8]	-	✓	✓	Person	211K	384 X 288
FLIR [9]	✓	✓	✓	Person, Cars, Poles, Bicycle, Bus, Dog	14K	640 X 512
KAIST [10]	✓	✓	✓	Person, Cars, Poles, Bicycle, Bus	95K	640 X 480
PTB-TIR [11]	✓	✓	✓	Person	30K	1280 X 720
CROSSIR [12]	✓	✓	✓	Person	14K	640 X 480

- A detailed comparative evaluation of SoA object detection based on a modified YOLO-v5 network, fine-tuned for thermal images using this newly acquired dataset.
- Model optimization using TensorRT inference accelerator to implement a fast inference network on SoA embedded GPU boards (Jetson, Xavier) with comparative evaluations.
- A determination of realistic frame rates that can be achieved for thermal object detection on SoA embedded GPU platforms.

II. BACKGROUND

ADAS (Advanced Driver Assistance Systems) are classified as AI-based intelligent systems integrated with core vehicular systems to assist the driver by providing a wide range of digital features for safe and reliable road journeys. Such type of system is designed by employing an array of electronic sensors and optical mixtures such as different types of cameras to identify surrounding impediments, driver faults, and reacts automatically.

The second part of this section will mainly summarize the existing/ published thermal datasets along with their respective attributes. These datasets can be effectively used for training and testing the machine learning algorithms for object detection in the thermal spectrum for ADAS. The complete dataset details are provided in Table I.

A. Related Literature

We can find numerous studies regarding the implementation of object detection algorithms using AI based conventional machine learning as well as deep learning algorithms. Such type of optical imaging-based systems system can be deployed and effectively used as forward sensing methods for ADAS. Advanced Driver-Assistance Systems (ADAS) is an active area of research that seeks to make road trips more safe and secure. Real-time object detection plays a critical role to warn the driver thus allowing them to make timely decisions [13]. Ziyatdinov *et al.* [13] proposed an automated system to detect road signs. This method uses the GTSRB dataset [14] to train on conventional machine learning algorithms which include SVM, KNN, and Decision Trees classifier. The results proved that SVM and K – nearest neighbour (k-NN) outperforms all other classifiers. Autonomous cars on the road require the ability to consistently perceive and comprehend their surroundings [15]. Oliver *et al.* [15] presented a procedure to use Bernoulli particle filter, which is suitable for object identification because it can handle a wide range of sensor measurements as well as object appearance-disappearance. Gang Yan *et al.* [16] proposed a novel method to use HOG to extract features and AdaBoost and SVM classifiers to detect vehicles in real-time. The histogram of oriented gradients (HOG) is a feature extraction technique used for object detection in the domain of computer vision and machine learning. The study concluded that the AdaBoost classification technique performed slightly better than SVM since it uses the ensemble method. Authors in [17], proposed another approach to detect vehicles on road using HOG filters to again extract features from the frames and then classify them using support vector machines and decision tree classification algorithms. Furthermore, SVM achieved 93.75% accuracy, which outperformed decision tree accuracy on classifying the vehicles. These are some of the conventional machine learning object detection techniques used for driver assistance system till date. The main drawback of traditional machine learning techniques is that the features are extracted and predefined prior to training and testing of the algorithms. When dealing with high-dimensional data, and with many classes conventional machine learning techniques are often ineffective [18].

Deep learning approaches have emerged as more reliable and effective solutions than these classic approaches. There are many state-of-the-art pre-trained deep learning classifiers and object detection models which can be retrained and rapidly deployed for designing efficient forward sensing algorithms [19]. YOLO (you only look once) object classifier provides sufficient performance to operate at real-time speeds on conventional video data without compromising the overall detector precision [20]. Veta *et al.* [21] presented a technique for detecting objects at a distance by employing YOLO on low-quality thermal images. Another research [22] focused on pedestrian detection in thermal images using the histogram of gradient (HOG) and YOLO methods on FLIR [9] dataset and computed performance with a 70% accuracy on test data using the intersection over union technique. Further, Rumi *et al.* [23] proposed a real-time human detection technique using YOLO-v3 on KAIST [10] thermal

TABLE II
COMPARISON ANALYSIS OF PREVIOUS YOLO VERSIONS WITH YOLO-V5

Yolo Version	Training Dataset	Validation mAP	FPS	Implementation Framework
YOLO [29]	VOC 2007 + 2012	63.4	45	DarkNet
YOLO-v2 (608x608) [30]	MSCOCO	48.1	40	DarkNet
YOLO-v3 (608x608) [30]	MS COCO	57.9	20	DarkNet
YOLO-v4 (608x608) [31]	MSCOCO	43.5	62	DarkNet
YOLO-v5 (640x640) [32] *X Large Model*	MSCOCO	68.9	83	PyTorch

dataset, achieving 95.5% average precision on test data. Authors in [24] proposed a human detection system using YOLO object detector. The authors used their custom dataset recorded in different weather conditions using FLIR Therma-CAM P10 thermal camera. Using five Siamese networks, the authors in [25] proposed a data-driven appearance score based on an innovative edge-based descriptor. The network was trained on a locally gathered single class pedestrians' dataset to obtain robust outcomes with an average precision of 86.2%.

Focusing on road-side objects, authors in [26] used YOLO-v2 object detection model to enhance the recognition of tiny vehicle objects by combining low-level and high-level features of the image. In [27], the authors proposed a deep learning-based vehicle occupancy detection system in a parking lot using a thermal camera. In this study authors had established that YOLO, Yolo-Conv, GoogleNet, and ResNet18 are computationally more efficient, take less processing time, and are suitable for real-time object detection. In one of the most recent studies [28], the efficacy of typical state-of-the-art object detectors which includes Faster R-CNN, SSD, Cascade R-CNN, and YOLO-v3 was assessed by retraining them on a thermal dataset. The results demonstrated that Yolo-v3 outclassed other object SoA object detectors. As compared to all the previous versions of YOLO released, YOLO-v5 has a Cross-Stage-Partial (CSP) backbone and PA-NET neck. The foremost improvements include mosaic data augmentation and auto learning bounding box anchors. The detailed comparative analysis of the recently released Yolo-v5 with all the previous versions is presented in Table II.

It can be observed from Table II that YOLO-V5 has comparatively achieved better validation results in terms of the highest mean average precision and frames per second on COCO dataset as compared to the previous version of YOLO framework. The optimal training and fine-tuning process of CNN to predict objects in low resolution, grayscale, and thermal infrared imaging (with lack of color information) and further optimizing the trained network to be deployed on edge devices is a challenging task [25]. For this task, Yolo-v5 open-source object detection framework is employed as it has better detection results than the previous YOLO versions as shown in Table II. In one of our recently published study [33], we have proposed a novel state-of-the-art YOLO-v5 based thermal object detection algorithm trained on public datasets and validated the performance on GPU with a maximum FPS rate of 170 and 3 FPS on Nvidia-Jetson Nano.

The main contribution of this research work is the establishment of a novel C3I thermal automotive dataset, which is then used to train the YOLO-v5 object detection algorithm along with four other public datasets. This study produced superior outcomes on thermal data and advances the state-of-the-art in the form of higher FPS rate and less inference time by optimizing the trained networks using TensorRT neural accelerator which were then deployed on both the edge-GPU devices which includes Nvidia Jetson Nano and Nvidia Jetson Xavier devices.

B. *Object Detection on Edge Devices*

AI on edge devices benefit us in various methods such that it speeds up decision-making, makes data processing more reliable, enhances user experience with hyper-personalization, and cuts down the costs. While machine learning models have shown immense strength in diversified consumer electronic applications, the increased prevalence of AI on edge has contributed to the growth of special-purpose embedded boards for various applications. Such types of embedded boards can achieve image inference at higher frames per second (fps) and low power usage. Some of these board includes Nvidia Jetson Nano, Nvidia Xavier, Google Coral, AWS DeepLens, and Intel AI-Stick. Authors in [34], [35] proposed a raspberry pi-based edge computing system to detect thermal objects. Sen Cao *et al.* [36] developed a roadside object detector using KITTI dataset [37] by training an efficient and lightweight neural network on Nvidia Jetson TX2 embedded GPU.

In another study [38] authors proposed deep learning-based smart task scheduling for self-driving vehicles. This task management module was implemented on multicore SoCs (Odroid Xu4 and Nvidia Jetson).

The overall goal of this study is to analyze the real-time performance feasibility of Thermal-YOLO object detector by deploying on edge devices. Different network variants of yolo-v5 framework are trained and fine-tuned on thermal image data and further deployed on the Nvidia Jetson Nano [2] and Nvidia Jetson Xavier NX [3]. These two platforms, although from the same manufacturer provide very different levels of performance and may be regarded as close to current SoA in terms of performance for embedded neural inference algorithms.

III. THERMAL DATA ACQUISITION AT SCALE FOR ADAS

This section will mainly cover the thermal data collection process using the LWIR prototype thermal imaging camera. The overall data is consisting of more than 35K distinct thermal frames acquired in different weather and environmental conditions. The data collection process includes shutterless camera calibration and thermal data processing [39], using the Lynred Display Kit (LDK) [40], data collection methods, and overall dataset attributes with different weather and environmental conditions for comprehensive data formation.

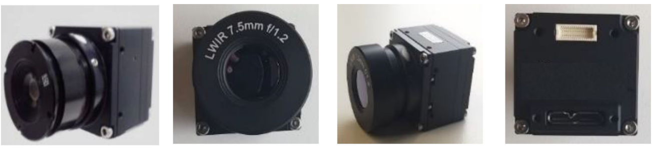


Fig. 1. LWIR thermal imaging module images from different view angles.

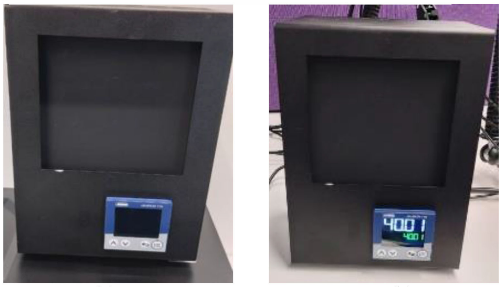
A. Prototype Thermal Camera

For the proposed research work we have utilized an uncooled thermal imaging camera developed under the HELIAUS project [1]. The main characteristic of this camera includes its low-cost, lightweight, and sleek compact design thus allowing to easily integrate it with artificially intelligent imaging pipelines for building effective in-cabin driver-passenger monitoring and road monitoring systems for ADAS. It enables us to capture high-quality thermal frames with low-power consumption thus proving the agility of configurations and data processing algorithms in real-time. Fig. 1 shows the prototype thermal camera. The technical specifications of the camera are as follows, the camera type is a VGA long-wave infrared (LWIR) with a spectral range from 8-14 µm and a camera resolution of 640x480 pixels. The focal length (f) of the camera is 7.5 mm, F-number is 1.2, the pixel pitch is 17 µm, and the power consumption is less than 950mW. The camera relates to a high-speed USB 3.0 (micro-USB) port for the interface. Moreover, the camera has a frame rate of 30 FPS. The camera has a thermal time constant of 12 ms. It is a time parameter that shows how quickly the bolometer reacts to the incoming flux change and reaches its expected level. Moreover, the camera comes with flat field correction (FFC) to remove non-uniformities in the thermal frames caused by optical factors. The FFC method nearly takes 100 ms to 300 ms time frame.

The data is recorded using a specifically designed toolbox. The complete camera calibration process along with the data processing pipeline is explained in the next section.

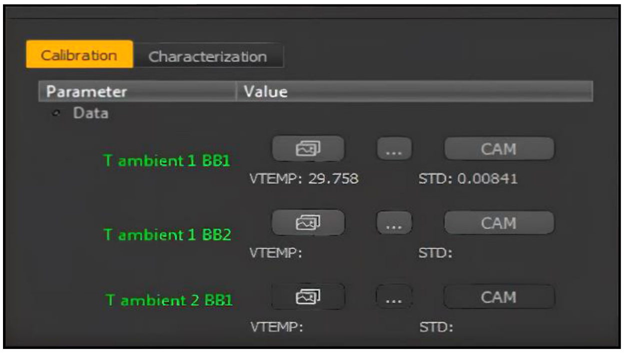
B. Shutterless Calibration and Real-Time Data Processing

This section will highlight the thermal camera calibration process for shutterless camera configuration along with real-time data processing methods for converting the raw thermal data to refined outputs. Shutterless technology allows uncooled IR engines and thermal imaging sensors to continuously operate without the need for a mechanical shutter for Non-Uniformity Correction (NUC) operations. Such type of technology provides proven and effective results in poor visibility conditions ensuring good quality thermal frames in real-time testing situations. For this, we have used a low-cost blackbody source to provide three different constant reference temperature values referred to as T-ambient1-BB1 (hot uniform scene with temperature value of 40-degree centigrade), T-ambient1-BB2 (cold uniform scene with the temperature value of 20 degrees centigrade), and T-ambient2-BB1 (either hot or cold uniform scene but with different temperature value). The imager can store up to 50 snapshots and select the best uniform temperature scenes for calibration purposes. Fig. 2 shows the blackbody used for the thermal camera calibration.



(a) (b)

Fig. 2. Thermal camera calibration (a) blackbody source used for LWIR thermal camera calibration, (b) uniform scene: temperature set to 40.01 degree centigrade.



Parameter	Value
Data	

T ambient 1 BB1 — VTEMP: 29.758 STD: 0.00841

T ambient 1 BB2 — VTEMP: STD:

T ambient 2 BB1 — VTEMP: STD:

Fig. 3. Prototype thermal camera SDK for loading constant reference temperatures values for shutterless camera calibration.

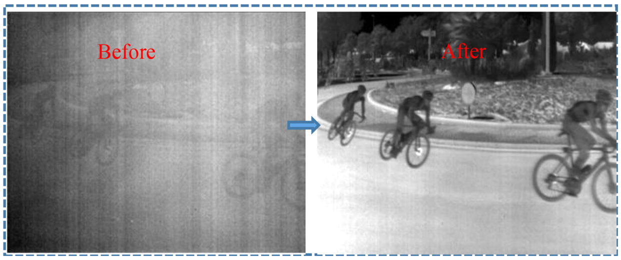


Fig. 4. Shutterless algorithm results on sample thermal frame captured from 640x480 LWIR thermal camera designed by Lynred France [40].

Once the uniform temperature images are recorded the images are loaded in camera SDK as shown in Fig. 3 to finally calibrate the shutterless camera stream. Fig. 4 shows the results before applying shutterless calibration and processed results using shutterless algorithms on thermal frame capture through the prototype thermal IR camera.

In the next phase, various real-time image processing-based correction methods are applied to convert the original thermal data to produce good-quality thermal frames. Fig. 5 shows the complete image processing pipeline.

As shown in Fig. 5 image processing pipeline consist of three different image correction methods which include gain

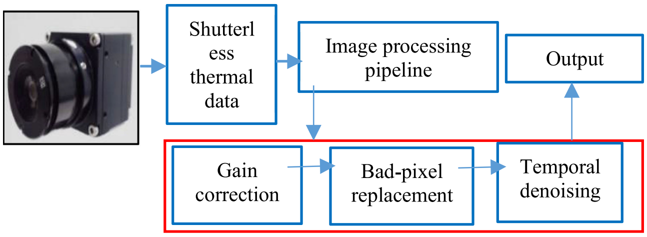


Fig. 5. Thermal image correction pipeline.

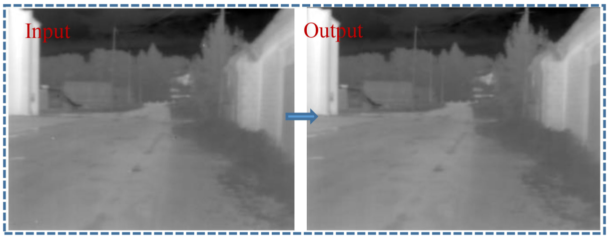


Fig. 6. Bad pixel replacement algorithm output on sample thermal frame, left side frame with some bad pixels and the right side is processed frame.

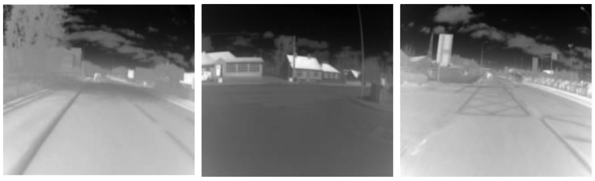


Fig. 7. High-quality thermal frames after applying the shutterless calibration algorithm and image correction methods.

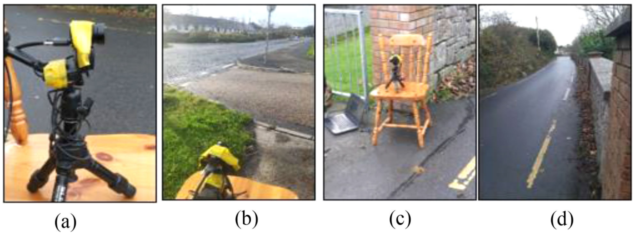


Fig. 8. Data Acquisition setup by placing the camera at a fixed place (a) camera mounted on a tripod stand, (b) complete daytime roadside view, (c) video recording setup at 30 fps, (d) evening time alleyway view.

correction, bad-pixel replacement, and temporal denoising. The further details of these methods are provided as follows.

1) Gain Correction Automatic Gain Control (AGC): Thermal image detectors, based on flat panels, suffer from irregular gains due to the non-uniform amplifiers. To correct the irregular gains, a common yet effective technique referred to as automatic gain control is applied. It is usually based on the gain map. By averaging uniformly illuminated images without any objects, the gain map is designed. By increasing the number of images for averaging provides a good gain-correction performance since the remained quantum noise in the gain map is reduced [40].

2) Bad Pixel Replacement (BPR): This is used to list bad pixels estimated at the calibration stage. It works by tracking potential new bad pixels by looking at pixel neighbourhood also known as the nearest neighbour method. Once it traces the bad pixels in the nearest neighbor it replaces them with good pixels. Fig. 6 demonstrates one such example.

3) Temporal Denoising (TD): The consistent reduction of image noise poses a frequently recurring problem in digitized thermal imaging systems and especially when it comes to un-cooled thermal imagers [41]. To mitigate these limitations for better outputs different methods are used which include hardware as well software-based image processing methods such as temporal and spatial denoising algorithms. The temporal denoising or temporal filtering method is typically performed to decrease the temporal noise and prevent temporal vibrations in the thermal frames. While acquiring the video sequence from an uncooled thermal camera, the pixel values can vary with the passage of time. This method is employed to smooth out the variations of pixel values at a given position thus producing refined thermal output. In commercial solutions, it usually works by gathering multiple frames and averaging those frames to cancel out the random noise among the frames. In our data acquisition process, this method is used after applying the shutterless algorithm. Fig. 7 shows the sample thermal images in the form of outcomes after applying shutterless algorithms and all the image processing-based corrections methods as shown in Fig. 5.

C. Data Collection Methods and Overall Dataset Attributes

This section will highlight different data collection approaches adopted in this research work. The data is collected in two different approaches. In, the first approach (M-1) the data is gathered in an immobile method by placing the camera at a fixed place. The camera is mounted on the tripod stand at a fixed height of nearly 30 inches such that the roadsides objects are covered in the video stream. The thermal video stream is recorded at 30 frames per second (FPS). The data is recorded in different weather and environmental conditions. Fig. 8 shows the M-1 data acquisition setup. In the second method (M-2) the thermal imaging system is mounted over the car and data is acquired in the mobile method. The prime reason for collecting the data in two different methods is to bring variations and collect distinctive local data in different environmental and weather conditions. For this, a specialized waterproof camera housing case was designed to hold the thermal camera in the correct position and angle to cover the entire roadside scene. The housing case is fixed on a suction-based tripod stand thus allowing us to easily fix and remove the complete structure from the car bonnet. The housing case also contains a visible camera to get initial visible images as reference data thus allowing us to adjust both the camera positions in proper angle and field of view.

Fig. 9 shows the camera housing case along with the initial data acquisition setup whereas

Fig. 10 shows the housing case fixed on the tripod structure and complete M-2 acquisition setup mounted on the car. The overall dataset is acquired from Galway County Ireland. The data is collected in form of short video clips and more than>35000

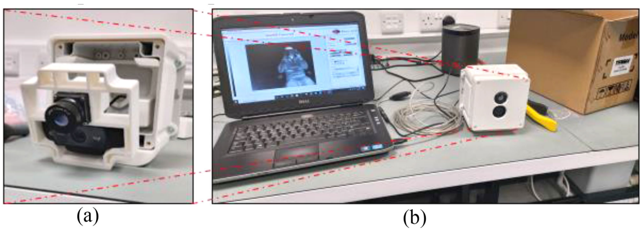


Fig. 9. Data acquisition setup through car (a) camera housing case holding thermal and visible camera, (b) initial data acquisition testing phase.

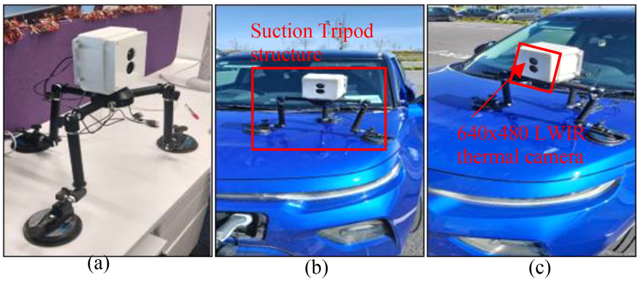


Fig. 10. Complete data acquisition setup mounted on the car (a) camera housing fixed on a suction tripod stand, (b) data acquisition kit from the front view, (c) data acquisition kit from the side view.

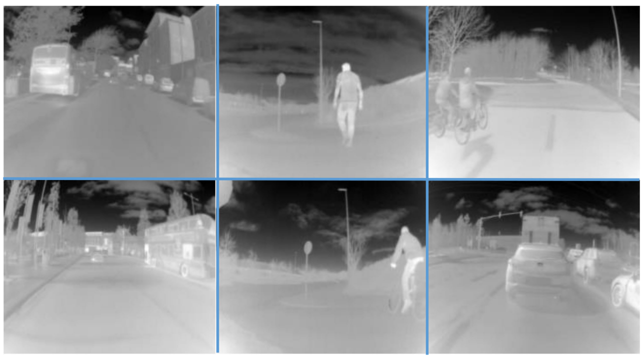


Fig. 11. Six different thermal samples acquired using LWIR 640 × 480 prototype thermal camera showing various class objects.

unique thermal frames have been extracted from the recorded video clips. The data is recorded in the daytime, evening time, and night-time which is distributed in the ratio of 44.61%, 31.78%, and 23.61% respectively of overall data. The complete dataset attributes are summarized in Table III. The acquired data comprises distinct stationary classes, such as road signs and poles, as well as moving object classes such as pedestrians, cars, buses, bikes, and bicycles.

Fig. 11 shows the six distinct sample of thermal frames captured in different environmental and weather conditions using M1 and M2 methods. These samples show different class objects such as buses, bicycles, poles, person, and cars. Most of these objects are found commonly on the roadside thus providing the driver a comprehensive video analysis of car surroundings.

TABLE III
New C3I Thermal Automative Dataset Attributes

Locally acquired dataset attributes				
Data collection method with frame properties	Total number of extracted frames	Processing Method	Environment	Time and weather conditions
M-1 Camera mounted at a fixed place 96 dpi (horizontal and vertical resolution) with 640x480 image dimension	8,140	Shutterless, AGC, BPR, TD	Roadside	Daytime with cloudy weather
	680		Alleyway	Evening time cloudy weather
	4,790		Roadside	Night-time with light cloudy and windy weather
M-2 Camera mounted on the car (Driving condition) 96 dpi (horizontal and vertical resolution) with 640x480 image dimension	9,600	Shutterless, AGC, BPR, TD	Industrial Park	Daytime with clear weather and light foggy weather
	11,960		Downtown	Evening time with partially cloudy and windy weather
	4,600	Shutterless, AGC, BPR, & TD	Downtown	Night-time with clear weather conditions
frames	*Daytime: 17,740 (44.61%)*	*Evening time: 12,640 (31.78%)*	*Night-time: 9,390 (23.61%)*	*Total: 39,770*

The recorded thermal datasets provide a greater number of thermal frames and extensive thermal data variations as compared to the FLIR open-source dataset. The acquired novel thermal datasets provide more than 35k distinct thermal frames. Moreover, the acquired LWIR dataset is collected in diverse weather, day, and environmental conditions which include daytime, evening time, and night-time with cloudy, windy, and light foggy weather conditions. Further, the newly proposed dataset is recorded in two different ways which include M1 and M2 methods where M1 refers to static data collection method by placing the camera at a fixed place and M2 refers to data collection method by mounting the camera on the car. The complete dataset attributes are provided in Table III.

IV. Proposed Methodology

This section will detail the proposed methodology and training outcomes from the various network variants tested in this study.

A. Network Training and Learning Perspectives

The overall training data comprises both locally and publicly available datasets. The complete training data is divided in the

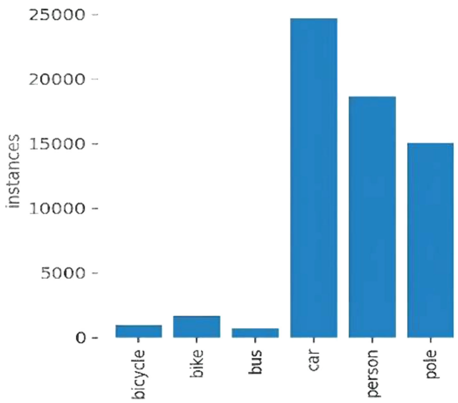


Fig. 12. Depicts the respective class-wise training samples distributions.

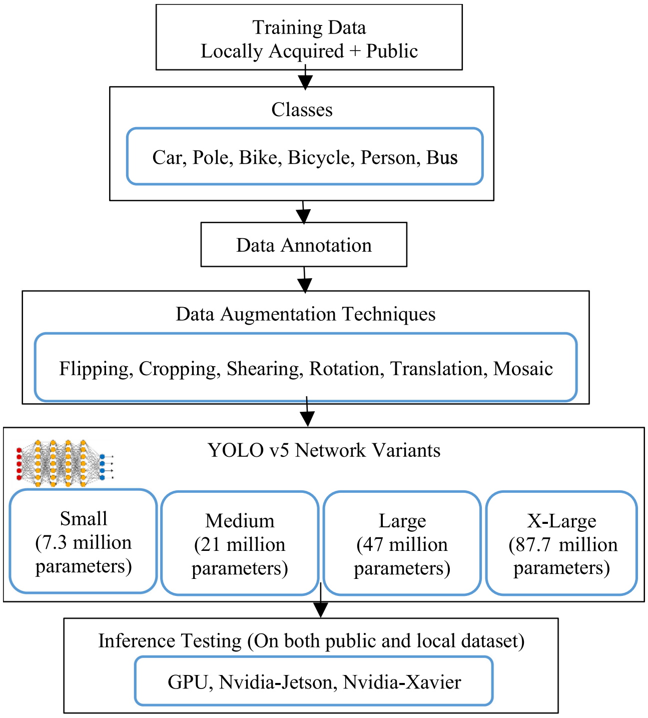

Training Data
Locally Acquired + Public
Classes
Car, Pole, Bike, Bicycle, Person, Bus
Data Annotation
Data Augmentation Techniques
Flipping, Cropping, Shearing, Rotation, Translation, Mosaic
YOLO v5 Network Variants
Small (7.3 million parameters)
Medium (21 million parameters)
Large (47 million parameters)
X-Large (87.7 million parameters)
Inference Testing (On both public and local dataset)
GPU, Nvidia-Jetson, Nvidia-Xavier

Fig. 13. Block diagram depicts the steps taken to evaluate the performance of YOLO v5 on local and public datasets.

ratio of 50% - 50% where 50% of data is selected from locally acquired thermal frames whereas the rest 50% of the training data is leveraged from public datasets. Six distinct types of roadside objects for driving assistance are included in training and validations sets. Fig. 12 shows the class-wise data distribution.

In the training phase of the YOLO-v5 [32] framework, a total of 59150 class-wise data samples were utilized, along with their corresponding class labels. Fig. 13 shows the complete block diagram representation of our algorithm to validate the performance of trained networks on the public as well as locally gathered datasets.

B. Data Annotation and Augmentation

The overall data annotations were performed manually using an open-source bounding box-based annotations tool LabelImg [42] for all the thermal classes in our study. Annotations are stored in YOLO format as text files. During the training phase all the YoloV5 network variations which include small, medium, large, and x-large networks have been trained to detect and classify six different classes in different environmental conditions.

Large-scale datasets are considered a vital requirement for achieving optimal training results using deep learning architectures. Without the need of gathering new data, data augmentation allows us to significantly improve the diversity of data available that can be effectively used for training the DNN models. In the proposed study we have incorporated a variety of data augmentation techniques which involve cropping, flipping, rotation, shearing, translation, mosaic transformation for an optimum training of all the network variants of the YOLO-v5 framework.

TABLE IV
TRAINING RESULTS

Optimizer: SGD (best model *)						
Network	P %	R%	mAP %	Box Loss	Object Loss	Classification Loss
Small	75.58	65.75	70.71	0.032	0.034	0.0017
Medium	71.06	64.74	65.34	0.027	0.030	0.0013
Large *	82.29	68.67	71.8	0.025	0.0287	0.0011
X-Large	74.23	65.03	64.94	0.025	0.0270	0.0010

C. Training Results

As discussed in Section A of Section IV all the networks are trained using the combination of public as well as the locally gathered dataset. Training data from public datasets are included from four different datasets which include FLIR [9], OST [4], CVC [5], and KAIST [10] datasets. Secondly, we have used thermal frames acquired from the locally gathered video sets using both M1 and M2 methods. The training process is performed on a server-grade machine with XEON E5-1650 v4 3.60 GHz processor, 64 GB of ram, and equipped with GEFORCE RTX 2080 Ti graphical processing unit. It comes with 12 GB of dedicated graphical memory, memory bandwidth of 616 GB/second, and 4352 cuda cores. During the training phase, the batch size is fixed to 32 and as an optimizer, both stochastic gradient descent (SGD) and ADAM optimizer were used. However, we were unable to achieve satisfactory training results using ADAM optimizer as compared to SGD thus selected SGD optimizer for training purposes. Table IV shows the performance evaluation of all the trained models in the form of mean average precision (mAP), recall rate, precision, and losses.

By analyzing Table IV, it can be observed that the large model performed significantly better when compared to other models with an overall precision of 82.29%, recall rate of 68.67%, and mean average precision of 71.8% mAP. Fig. 14 shows the graph results of yolo-v5 large model. The figure visualizes obtained PR-curve, box loss, object loss, and classification loss. During the training process, the X-large model consumes the maximum amount of hardware resources with the largest training time as

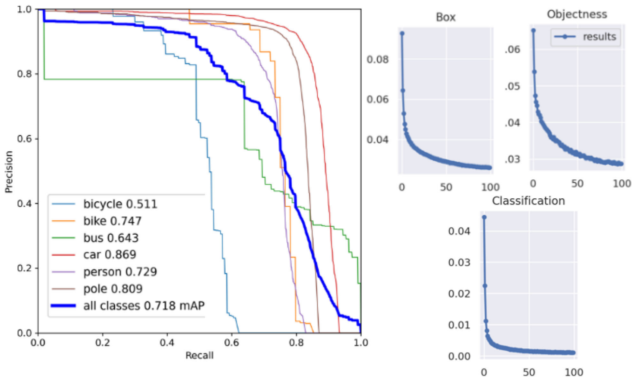
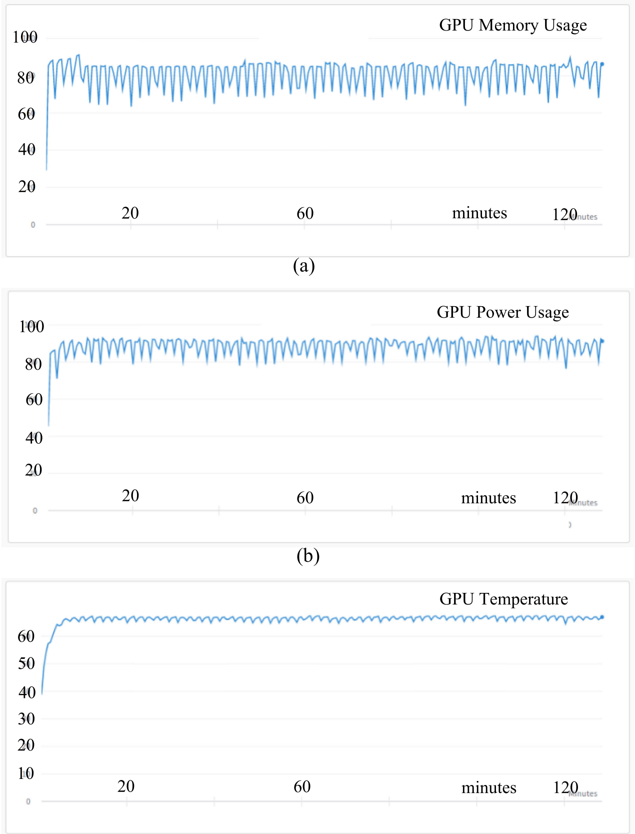


Fig. 14. Training results of YOLO-v5 large model using SGD optimizer.



(a)

(b)

(c)

Fig. 15. GPU resource utilization during the training process of x-large network, (a) 85% (9.78 GB) of GPU memory utilized, (b) 90% (585 watts) of GPU power required and, (c) 68 C of GPU temperature with the maximum rating of 89 C.

compared to other network variants with overall GPU usage of 9.78 GB and a total training time of 14 hours. Fig. 15 shows the overall GPU memory usage, GPU power required in percentages, and GPU temperature in centigrade scale while training the largest x-large network variant of the yolo-v5 framework.

V. Validation Results on GPU and Edge Devices

This section will demonstrate the object detection validation results on GPU as well as on two different embedded boards.

TABLE V
Test Dataset

Test Dataset Attributes					
Frames Used					
Public dataset	OST	CVC-09	KAI ST	FLIR	Total No frames
	50	5360 (day + night-time)	149	130	5,689
Local dataset	Method (M1)		Method (M2)		Total No frames
	8,820		16,560		25,380
					Total: 31,069

A. *Testing Methodology and Overall Test Data*

In this research study, we have used three different testing approaches which include the conventional test-time method with no augmentation (NA), test-time augmentation (TTA), and test-time with model ensembling (ME). TTA is an extensive application of data augmentation applied to the test dataset. It performs by creating multiple augmented copies of each image in the test set, having the model make a prediction for each, then returning an ensemble of those predictions. However, since the test dataset is enlarged with a new set of augmented images the overall inference time also increases as compared to NA which is one of the downsides of this approach. TTME or ensemble learning refers to as using multiple trained networks at the same time in a parallel manner to produce one optimal predictive inference model [43]. In this study, we have tested the performance of individually trained variants of the Yolo-v5 framework and selected the best combination of models which in turn helps in achieving better validation results.

After training all the networks variants of yolo-v5, the performance of each model is cross-validated on a comprehensive set of test data selected from the public as well as locally gathered novel thermal data. Table V provides the numeric data distribution of the overall validation set.

B. *Inference Results Using YOLO Network Variants*

In the first phase, we have run the rigorous inference test on GPU as well as Edge-GPU platforms on our test data using the newly trained networks variants of yolo framework. The overall test data is consisting of nearly ≈31000 thermal frames. Fig. 16 shows the inference results on 9 different thermal frames selected from both public as well as locally acquired data. These frames have data complications such as multiple class objects, occlusion, overlapping classes, scale variation, and varying environmental conditions. The complete inference results are available on our github repository (https://github.com/MaliFarooq/Thermal-YOLO).

In the second phase, we have run the combination of different models in a parallel manner using the model ensembling approach to output one optimal predictive engine which can be further used to run the inference test on the validation set.

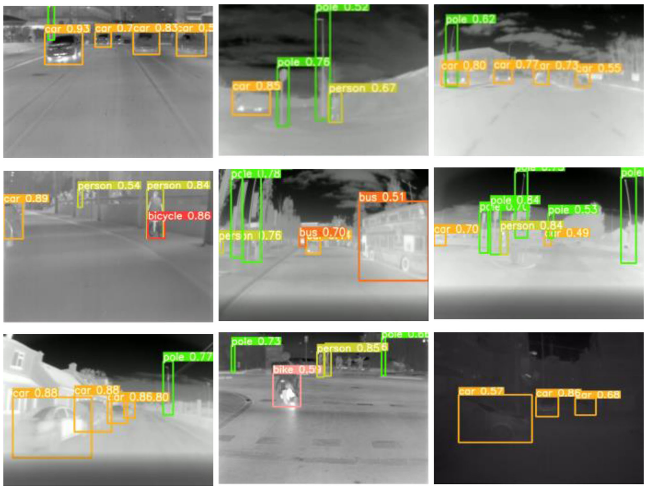


Fig. 16. Inference results on nine different frames selected from test data.

TABLE VI
MODEL ENSEMBLING

Model Combinations					
No	Small	Medium	Large	X-Large	Combination
State 1 (active) or 0 (not active)					
1	*1*	*1*	*0*	*0*	A0
2	*1*	*0*	*1*	*0*	A1
3	*1*	*0*	*0*	*1*	A2
4	*0*	*1*	*1*	*0*	A3
5	*0*	*0*	*1*	*1*	A4

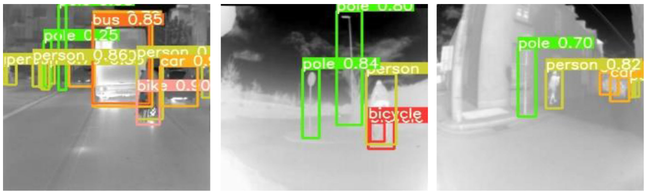


Fig. 17. Inference results on three different frames using model ensembling method.

The different combination of these models is shown in Table VI respectively where 1 indicates that the model is in active state and 0 means the model is in a non-active state.

With the model ensembling method small and large models (A1) turn out to best model combination in terms of achieving the best mAP, recall, and relatively less amount of inference time per frame thus producing optimal validation results. These results are examined in further parts of this section. Fig. 17 shows the inference results using A1 model ensembling engine on three different thermal frames selected from the test data. The first frame is selected from the public dataset whereas the other two frames are selected from the locally acquired thermal dataset. By closely analyzing the results it can be observed that model ensembling based inference engine has performed significantly well on diversified test data with image complexities like occlusions and overlapping classes.

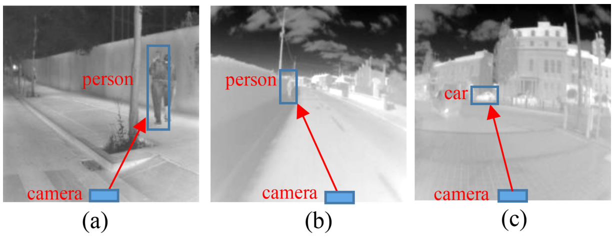
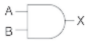


Fig. 18. Test data samples with the object at varying distances from the camera, (a) near-field distance, (b) mid-field distance, (c) far-field distance.

TABLE VII
QUANTITATIVE RESULTS ON GPU

Platform: GPU Inference image size: 800 x 800								
Confidence Threshold: 0.4, IoU Threshold: 0.6								
No Augmentation (NA)					Test-time Augmentation (TTA)			
Network	P %	R %	mAP%	FPS	P %	R %	mAP%	FPS
Small	72	46	43	79	76	48	50	45
Medium	73	54	49	53	76	58	57	26
Large	75	56	52	34	77	63	60	16
X-Large	74	53	49	20	71	59	55	10
Confidence Threshold: 0.2, IoU Threshold: 0.4								
NA					TTA			
Small	66	50	47	82	64	55	52	45
Medium	66	57	51	53	77	58	59	27
Large	71	61	56	35	78	63	63	16
X-Large	70	54	50	21	68	62	56	10
Confidence Threshold: 0.1, IoU Threshold: 0.2								
NA					TTA			
Small	65	52	48	81	65	53	53	45
Medium	69	54	51	53	77	58	59	26
Large	73	61	57	34	79	63	63	16
X-Large	71	54	52	21	69	62	57	10
Confidence Threshold: 0.2, IoU Threshold: 0.4								
Model Ensembling (ME)								
A = Small B = Large Comb: A1 	---	---	---	---	77	66	65	25

C. *Quantitative Validation Results on GPU*

The third part of the testing phase shows the quantitative numerical results of all the trained models on GPU. To better analyze and validate the overall performance for all the trained models on test data, relatively a smaller set of test images has been selected from the overall test set. For this purpose, a subset of 402 thermal frames is selected to compute all the evaluation metrics. The selected images consist of different roadside objects such as pedestrians, cars, and buses under different illumination and environmental conditions, time of day, and distance from the camera. The objects are either far-field (between 11-18 meters), mid-field (between 7-10 meters), or near-field (between 3-6 meters) from the camera. Fig. 18 shows selected views from the test data for quick reference of the reader.

The performance evaluation of each model is computed using four different metrics which include recall, precision, mean average precision (mAP), and frames per second rate (FPS). Table VII shows all the quantitative validation results on GPU.

TABLE VIII
CLASS-WISE QUANTITATIVE RESULTS

Class	Small model	Medium model	Large model	X-large model
	Average Precesion%			
Bicycle	18.9	38.2	63.3	16.8
Bike	36.0	39.1	50.4	59.1
Bus	32.8	23.1	29.0	33.8
Car	76.2	79.4	77.0	74.8
Person	72.5	72.8	72.7	73.2
Pole	52.9	54.1	51.8	52.5
All classes	mAP = 48.2 ≈ 48	mAP = 51.1 ≈ 51	mAP = 57.4 ≈ 57	mAP = 51.7 ≈ 52

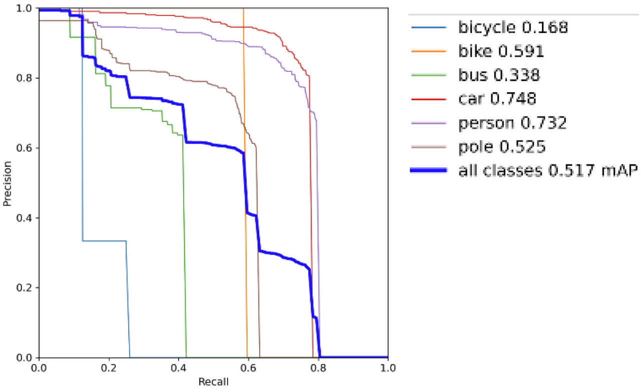


Fig. 19. Precision-Recall (PR) cruve of X-Large model showing the average precision value for all the six classes.

During the testing phase batch size is fixed to 8. Also, three different testing configuration is selected thus having separate confidence threshold values and the intersection of union values at each validation phase. Confidence threshold defines the minimum threshold value, or in other words, it is the minimum confidence score above which we consider a prediction as true. If it's below the threshold value, we consider the prediction as "no". The last row of Table VII shows the best ME results using A1 configuration from Table VI with a selected confidence threshold of 0.2 and IoU threshold of 0.4.

For futher in-depth analysis of all the trained network variants of YOLO-v5 framework we have presented class-wise quantitative results. The Table VIII shows the individual average precision, for all the six classes of four different thermally tuned models. It should be noted that these results are extracted using test time with no augmentation approach with confidence threshold of 0.1 and IoU threshold of 0.2.

As it can be observed from above Table VIII that although the large model has achieved the highest mean average precision however by observing the class-wise performance, the X-large model has achieved the highest average precision value for the maximum number of classes (i.e. bike, bus, and person). Fig. 19 shows the precision-recall curve for all the classes of the x-large model.

D. *Quantitative Validation Results on Edge-GPU Devices*

This section will review the quantitative validation results on two different Edge-GPU platforms (Jetson Nano & Jetson Xavier NX). It is pertinent to mention that Jetson Xavier NX development kit embeds more computational power in terms of GPU, CPU, and memory as compared to Nvidia Jetson Nano. Table IX shows the hardware specification comparison of both boards.

TABLE IX
HARDWARE SPECIFICATION COMPARISON

Hardware specification comparison of Nvidia Jetson Nano and Nvidia Jetson Xavier NX		
Board	Jetson Nano [23]	Jetson Xavier NX [25]
CPU	Quad-Core ARM® Cortex® -A57 MPCore, 2 MB L2, Maximum Operating Frequency: 1.43 GHz	6-core NVIDIA Carmel ARM®v8.2 64-bit CPU, 6 MB L2 + 4 MB L3, Maximum Operating Frequency: 1.9 GHz
GPU	128-core Maxwell GPU, 512 GFLOPS (FP16), Maximum Operating Frequency: 921 MHz	384 CUDA® cores + 48 Tensor cores Volta GPU, 21 TOPS, Maximum Operating Frequency: 1100 MHz
RAM	4 GB 64-bit LPDDR4 @ 1600MHz | 25.6 GB/s	8 GB 128-bit LPDDR4x @ 1600MHz | 51.2GB/s
On module Storage	16 GB eMMC 5.1 Flash Storage, Bus Width: 8-bit, Maximum Bus Frequency: 200 MHz (HS400)	
Thermal Design Power	5W – 10W	10W – 15W
AI Performance	0.5 TFLOPS (FP16)	6 TFLOPS (FP16) 21 TOPS (INT8)

On Jetson Nano we have validated the performance of the small version only whereas on Jetson Xavier NX we have evaluated the performance of smaller and medium versions of models due to the memory limitations and constrained hardware resources on these boards. During the testing phase, we have selected the highest power modes on both boards to provide the utmost efficiency thus utilizing maximum hardware resources. For instance, on Nvidia Xavier board NX we have selected 'Mode Id: 2' which means the board is operating in 15-watt power mode with all the six cores active with a maximal CPU frequency of 1.4 gigahertz and GPU frequency of 1.1 gigahertz. Similarly, on Nvidia Jetson Nano all the four CPU cores were utilized with overall power utilization of 5 watts. Table X shows the quantitative validation results on ARM processor based embedded boards.

For a better and more comprehensive valuation of thermally tuned networks, we have demonstrated the performance of the smaller network variant on various environmental conditions as shown in Table XI which includes alleyways, roadside, industrial park, and downtown. The main reason for shortlisting the smaller network variant is it requires the least computational resources and shows effective results on both edge-GPU devices. As it can be observed from Table XI we have obtained the highest mean average precision of 71.9% on roadside environmental thermal frames which is highlighted in green color.

E. *Real-Time Hardware Feasibility Testing*

While running these tests we closely monitor the temperature ratings of different hardware peripherals on both Edge-GPU

TABLE X
QUANTITATIVE RESULTS ON EDGE PLATFORMS

Platform: Nvidia Jetson Nano Inference image size: 128 x 128								
Confidence Threshold: 0.4, IoU Threshold: 0.6								
	NA				TTA			
	P %	R %	mA P%	FPS	P %	R %	mA P%	FPS
Small	75	44	45	3	77	47	49	1
Confidence Threshold: 0.2, IoU Threshold: 0.4								
	NA				TTA			
Small	75	44	47	3	71	51	51	1
Confidence Threshold: 0.1, IoU Threshold: 0.2								
	NA				TTA			
Small	66	47	48	2	73	50	52	1
Platform: Nvidia Jetson Xavier NX Inference image size: 128 x 128								
Confidence Threshold: 0.4, IoU Threshold: 0.6								
	NA				TTA			
Small	75	44	45	18	77	47	49	10
Med	76	53	50	12	79	50	52	6
Confidence Threshold: 0.2, IoU Threshold: 0.4								
	NA				TTA			
Small	75	44	47	19	71	51	51	10
Med	76	52	53	12	73	54	53	6
Confidence Threshold: 0.1, IoU Threshold: 0.2								
	NA				TTA			
Small	66	47	48	18	73	50	52	10
Med	76	51	52	12	81	49	53	6

TABLE XI
QUANTITATIVE RESULTS ON DIFFERENT ENVIRONMENTAL CONDITIONS

Platform: Nvidia Jetson Nano and Nvidia Jetson Xavier Inference image size: 128 x 128 (best results*)			
Small network variant with Confidence Threshold: 0.1, IoU Threshold: 0.2			
Environmental Conditions	mAP %	Precision %	Recall %
Alleyway	56.5	66.4	63.7
Roadside*	71.9	74.6	69.2
Industrial park	51.2	71.2	57
Downtown	38.1	59.8	38.7

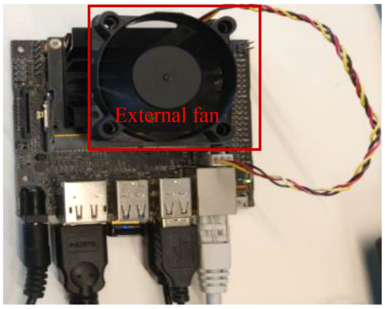
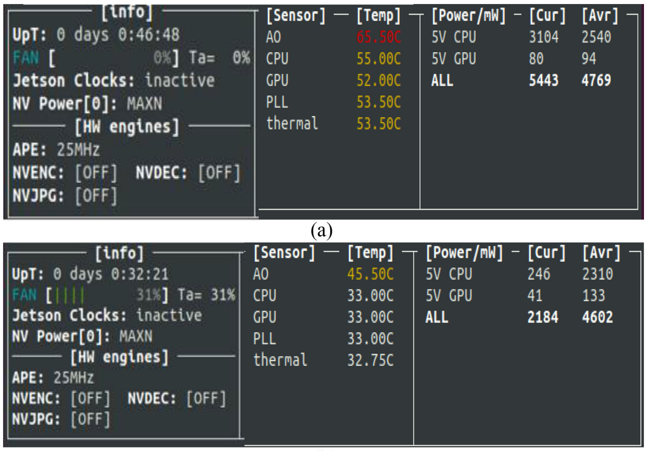


Fig. 20. External 5-volt fan unit mounted on Nvidia Jetson Nano processor heatsink to avoid onboard overheating effect while running the inference testing.



```
──────── [info] ─────────  [Sensor] ─ [Temp] ┬ [Power/mW] - [Cur]  [Avr] ─
UpT: 0 days 0:46:48        AO         65.50C │ 5V CPU       3104   2540
FAN [             0%] Ta=  0%   CPU   55.00C │ 5V GPU       80     94
Jetson Clocks: inactive    GPU        52.00C │ ALL          5443   4769
NV Power[0]: MAXN          PLL        53.50C │
────── [HW engines] ─────  thermal    53.50C │
APE: 25MHz                                   │
NVENC: [OFF]  NVDEC: [OFF]                   │
NVJPG: [OFF]                                 │
```

(a)

```
──────── [info] ─────────  [Sensor] ─ [Temp] ┬ [Power/mW] - [Cur]  [Avr] ─
UpT: 0 days 0:32:21        AO         45.50C │ 5V CPU       246    2310
FAN [||||       31%] Ta= 31%  CPU     33.00C │ 5V GPU       41     133
Jetson Clocks: inactive    GPU        33.00C │ ALL          2184   4602
NV Power[0]: MAXN          PLL        33.00C │
────── [HW engines] ─────  thermal    32.75C │
APE: 25MHz                                   │
NVENC: [OFF]  NVDEC: [OFF]                   │
NVJPG: [OFF]                                 │
```

(b)

Fig. 21. Temperature rating difference of different onboard hardware peripherals on Jetson Nano (a) without fan: A0 thermal zone = 65.50 C, CPU = 55 C, GPU = 52 C, PLL: 53.50, overall thermal temperature = 53.50 C, (b) with external fan: A0 thermal zone = 45.50 C, CPU = 33 C, GPU = 33 C, PLL: 33, overall thermal temperature = 32.75 C.

platforms. It is done to prevent the overheating effect which can damage the onboard processor or effect the overall operational capability of the system. In the case of Nvidia Jetson Nano, a cooling fan was mounted on top of the processor heatsink to reduce the overheating effect as shown in Fig. 20.

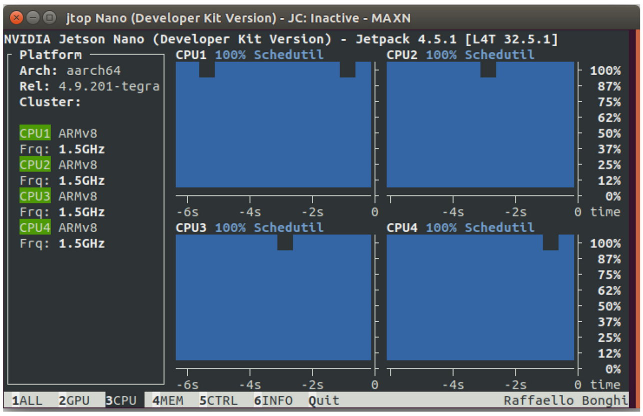
The temperature ratings of various hardware peripherals are monitored using eight different on-die thermal sensors and one on-die thermal diode. These temperature monitors are referred to as CPU-Thermal, GPU-Thermal, Memory-Thermal, and PLL-Thermal (part thermal zone). External fans help us in reducing the temperature rating of various hardware peripherals drastically as compared to without mounting the fan. For this jetson-stats open-source python library [44] have been used. Jetson-stats is a package for monitoring various onboard hardware resources such as real-time information of CPUs status, Memory, GPU, disk, fan, temperature rating, and all status about Jetson clocks. Fig. 21 shows the temperature rating difference of onboard thermal sensors while running the smaller version of the model on Nvidia Jetson Nano without and with mounting the external cooling fan.

It can be examined from Fig. 21(b) that by mounting an external cooling fan the temperature rating of various onboard peripheral on Jetson Nano was reduced by nearly 30% thus allowing us to operate the board at its maximum capacity for rigorous model testing. Fig. 22 shows the Nvidia Jetson running at its full pace (with an external fan) such that all the four cores running at their maximum limit (100% capacity) while running the quantitative and inference test by deploying the smaller network variant of the yolo-v5 framework.

Fig. 23 shows the temperature rating difference of onboard thermal sensors while running the smaller version of the model on Nvidia Jetson Xavier NX board. Whereas Fig. 24 shows the CPU and GPU usage while running the smaller variant of YOLO-v5 framework for quantitative validation and inference test on Nvidia Xavier NX development kit.

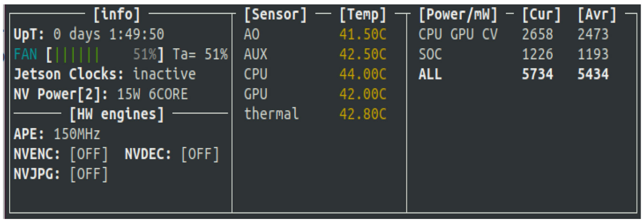
VI. MODEL PERFORMANCE OPTIMIZATION(S)

This section will mainly aim at further model optimization using TensorRT [45] inference accelerator tool. The prime reason for this is to further increase the FPS rate for real-time evaluation and on-board feasibility testing on edge devices. Secondly, it



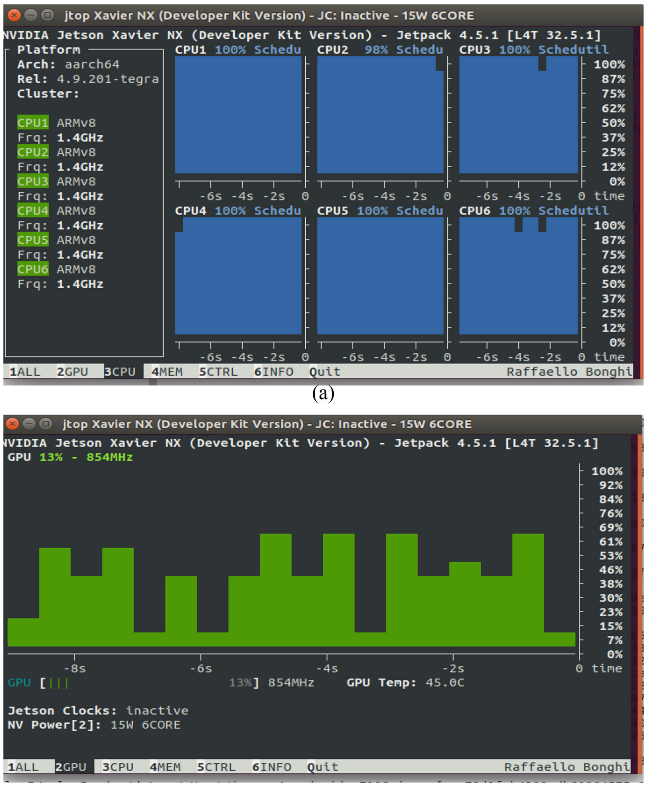
```
jtop Nano (Developer Kit Version) - JC: Inactive - MAXN
NVIDIA Jetson Nano (Developer Kit Version) - Jetpack 4.5.1 [L4T 32.5.1]
 Platform          CPU1 100% Schedutil        CPU2 100% Schedutil
 Arch: aarch64
 Rel: 4.9.201-tegra
 Cluster:

 CPU1 ARMv8
 Frq: 1.5GHz
 CPU2 ARMv8
 Frq: 1.5GHz
 CPU3 ARMv8
 Frq: 1.5GHz
 CPU4 ARMv8
 Frq: 1.5GHz
                   CPU3 100% Schedutil        CPU4 100% Schedutil

 1ALL  2GPU  3CPU  4MEM  5CTRL  6INFO  Quit            Raffaello Bonghi
```

Fig. 22. Nvidia Jetson Nano running at MAXN power mode with all the cores running at their maximum capacity while running the inference test and quantitative validation test.



```
───────── [info] ─────────┬ [Sensor] ─ [Temp] ─┬ [Power/mW] ─ [Cur]  [Avr] ─
UpT: 0 days 1:49:50       │ AO         41.50C  │ CPU GPU CV    2658   2473
FAN [||||||     51%] Ta= 51%│ AUX      42.50C  │ SOC           1226   1193
Jetson Clocks: inactive   │ CPU        44.00C  │ ALL           5734   5434
NV Power[2]: 15W 6CORE    │ GPU        42.00C  │
───── [HW engines] ─────  │ thermal    42.80C  │
APE: 150MHz               │                    │
NVENC: [OFF]  NVDEC: [OFF]│                    │
NVJPG: [OFF]              │                    │
```

Fig. 23. Temperature rating of different onboard hardware peripherals on Jetson Xavier NX (a) A0 thermal zone = 41.50 C, AUX: 42.5 C, CPU = 44 C, GPU = 42 C, overall thermal temperature = 42.80 C.



```
jtop Xavier NX (Developer Kit Version) - JC: Inactive - 15W 6CORE
NVIDIA Jetson Xavier NX (Developer Kit Version) - Jetpack 4.5.1 [L4T 32.5.1]
 Platform        CPU1 100% Schedu  CPU2  98% Schedu  CPU3 100% Schedutil
 Arch: aarch64
 Rel: 4.9.201-tegra
 Cluster:

 CPU1 ARMv8
 Frq: 1.4GHz
 CPU2 ARMv8
 Frq: 1.4GHz
 CPU3 ARMv8
 Frq: 1.4GHz
 CPU4 ARMv8
 Frq: 1.4GHz
 CPU5 ARMv8
 Frq: 1.4GHz
 CPU6 ARMv8
 Frq: 1.4GHz
                 CPU4 100% Schedu  CPU5 100% Schedu  CPU6 100% Schedutil

 1ALL  2GPU  3CPU  4MEM  5CTRL  6INFO  Quit            Raffaello Bonghi
```

(a)

```
jtop Xavier NX (Developer Kit Version) - JC: Inactive - 15W 6CORE
NVIDIA Jetson Xavier NX (Developer Kit Version) - Jetpack 4.5.1 [L4T 32.5.1]
 GPU 13% - 854MHz

 GPU [|||                  13%] 854MHz     GPU Temp: 45.0C

 Jetson Clocks: inactive
 NV Power[2]: 15W 6CORE

 1ALL  2GPU  3CPU  4MEM  5CTRL  6INFO  Quit            Raffaello Bonghi
```

(b)

Fig. 24. Nvidia Jetson Xavier running at 15-watt 6 core power mode, (a) all the CPU cores running at its maximum capacity while running the quantitative validation test, (b) 69% GPU utilization while running the inference test with an image size of 128 x 128.

TABLE XII
TensorRT Inference Accelerator Results

FPS on Nvidia Jetson Nano and Nvidia Jetson Xavier NX		
Board	Nvidia Jetson Nano	Nvidia Jetson Xavier NX
Test Data	402 images with the resolution of 128x128	
Overall inference time	35,090 milliseconds ≈ 35.1 seconds	6,675 milliseconds ≈ 6.7 seconds
PS	35.1 sec / 402 frames = 0.087 sec/frame FPS: 1 sec / 0.087 = 11.49 ≈ 11 fps	6.7 sec / 402 frames = 0.0166 sec/frame FPS: 1 sec / 0.0166 = 60.24 ≈ 60 fps

helps in saving onboard memory footprints on the target device by performing various optimization methods.

TensorRT [45] works by performing five modes of optimization methods for increasing the throughput of deep neural networks. In the first step, it maximizes throughput by quantizing models to 8-bit integer data type or FP16 precision while preserving the model accuracy. This method significantly reduces the model size since it is transformed from originally FP32 to FP16 version. In the next step, it uses layer and tensor fusion techniques to further optimize the usage of onboard GPU memory. The third step includes performing kernel auto-tuning. It is the most important step where the TensorRT engine shortlists the best network layers, and optimal batch size based on the target GPU hardware. In the second last step, it minimizes memory footprints and re-uses memory by distributing memory to tensor only for the period of its usage. In the last steps, it processes multiple input streams in parallel and finally optimizes neural networks periodically with dynamically generated kernels [45].

In the proposed research work we have deployed a smaller variant of yolo-v5 using TensorRT inference accelerator on both edge platforms Nvidia Jetson Nano and Nvidia Jetson Xavier NX development boards to further excel the performance of the trained model. It produces faster inference time thus increasing the FPS on thermal data which in turn helps us in building an effective real-time forward sensing system for ADAS embedded applications. Fig. 25 depicts the block diagram representation of deployment phase TensorRT inference accelerator on embedded platforms. Table XII shows the overall inference time along with FPS rate on thermal test data using TensorRT run-time engine. By analyzing the results from Table XII , we can deduce that TensorRT API supports in boosting the overall FPS rate on ARM-based embedded platforms by nearly 3.5 times as compared to the FPS rate achieved by running the non-optimized smaller variant on Nvidia Jetson Nano and Nvidia Jetson Xavier boards. The same is demonstrated via graphical chart results in Fig. 26.

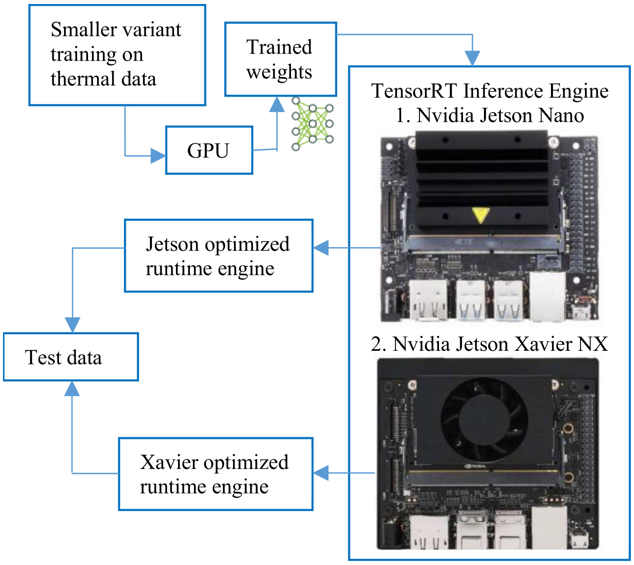

Smaller variant training on thermal data
Trained weights
GPU
TensorRT Inference Engine
1. Nvidia Jetson Nano
Jetson optimized runtime engine
Test data
2. Nvidia Jetson Xavier NX
Xavier optimized runtime engine

Fig. 25. Overall block diagram representation of deployment and running TensorRT inference accelerator on two different embedded platforms.

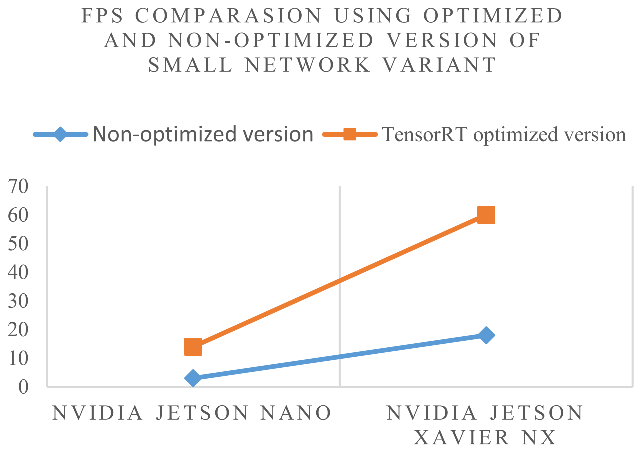


Fig. 26. FPS increment rate of nearly 3.5 times on Jetson Nano and Jetson Xavier NX embedded boards using the TensorRT built optimized inference engine.

Fig. 27 shows the thermal object detection inference results on six different thermal frames from the public as well as locally acquired test data produced through the neural accelerator.

VII. Discussion/Analysis

This section will review the training and testing performance of all YOLO-v5 framework model variants.

- During the training phase, the larger network variant of YOLO-v5 outperforms other network variants scoring the highest precision of 82.29% and a mean average precision (mAP) score of 71.8%.
- Although the large network variant performed significantly better during the training phase, the small network variant also performed well with an overall precision of 75.58% and mAP of 70.71%. Also, it gains a higher FPS rate on GPU during the testing phase as compared to the large model. Fig. 28 summarizes the quantitative performance comparison of small and large network variants of yolo framework.
- Due to the lesser number of model parameters of smaller architecture as compared to larger network variant (7.3M Vs 47M model parameters) and faster FPS rate on GPU during the testing phase as shown in Fig. 27 this model is shortlisted for validation and deployment purposes on both the edge embedded platforms Nvidia Jetson Nano and Nvidia Jetson Xavier NX kits.
- During the testing phase, it was noticed that by reducing the confidence threshold from 0.4 to 0.1 and the IoU threshold from 0.6 to 0.2 in three stepwise intervals, the model's mAP and recall rates increased significantly, but the precision level decreases. However, the FPS rate remains effectively constant in most of the trained model cases.
- TTA methods achieved improved testing results when compared to the NA method however the main drawback of this method is that the FPS rate drops substantially which is not suitable for real-time deployments. To overcome this problem a model ensembling (ME) based inference

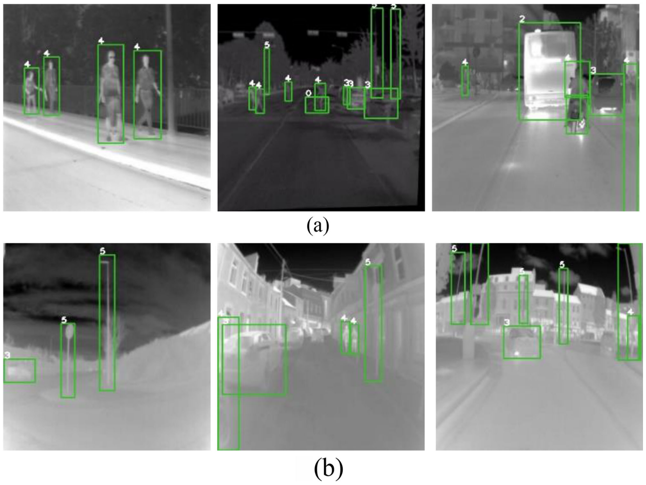


Fig. 27. Inference results using optimized smaller variant through TensorRT neural accelerator, (a) Object detection results on public data, (b) Object Detection results on locally acquired thermal frames.

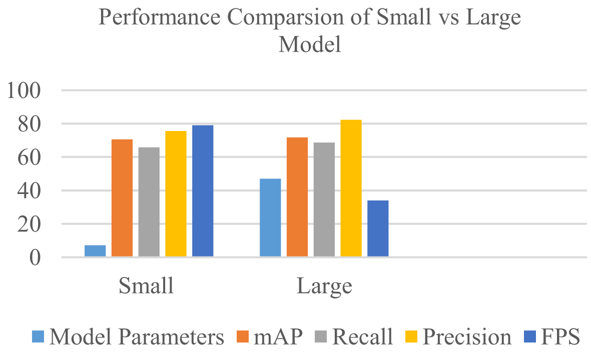


Fig. 28. Quantitative metrics comparison of small and large network variants.

engine is proposed. Table VII shows the ME results by running large & small model in parallel configuration with a confidence threshold of 0.2, and an IoU Threshold of 0.4. The ensembling engine attains an overall mAP of 66% with a frame rate of 25 FPS.

- When comparing the individual hardware resources of both the edge platforms (NVidia Jetson Nano and Jetson Xavier), Xavier is computationally more powerful than the Jetson Nano. Note that due to memory limitations and the lower computational power of the Jetson only the small network variant was evaluated on the Jetson Nano, whereas both the smaller and medium network variants were evaluated on the Jetson Xavier NX.
- It was observed that throughout the testing phase, it was important to keep a close eye on the operational temperature ratings of different onboard thermal sensors to avoid overheating, which might damage the onboard components or affect the system's typical operational performance. Active cooling fans were used on both boards during testing, and both ran at close to their rated temperature limits.
- This study also included model optimization using TensorRT [45] inference accelerator tool. It was determined that TensorRT leads to an approximate increase of FPS rate by a factor of 3.5 when compared to the non-optimized smaller variant of yolo-v5 on Nvidia Jetson Nano and Nvidia Jetson Xavier devices.
- After performing model optimization, the Nvidia Jetson produced 11 FPS and Nvidia Jetson Xavier achieved 60 FPS on test data.

VIII. Conclusion

Thermal imaging provides superior and effective results in challenging environments such that in low lighting scenarios and has aggregate immunity to visual limitations thus making it an optimal solution for intelligent and safer vehicular systems**.** In this study, we presented a new benchmark C3I thermal automotive dataset that comprises over 35K distinct frames recorded, analyzed, and open-sourced in challenging weather and environmental conditions utilizing a low-cost yet reliable uncooled LWIR thermal camera. All the YOLO v5 network variants were trained using locally gathered data as well as four different publicly available datasets. The performance of trained networks is analyzed on both GPU as well as ARM processor-based edge devices for onboard automotive sensor suite feasibility testing. On edge devices, the small and medium network edition of YOLO is deployed and tested due to certain memory limitations and less computational power of these boards. Lastly, we further optimized the smaller network variant using TensorRT inference accelerator to explicitly increase the FPS on edge devices. This allowed the system to achieve 11 frames per second on jetson nano, while the Nvidia Jetson Xavier delivered a significantly higher performance of 60 frames per second. These results validate the potential for thermal imaging as a core component of ADAS systems for intelligent vehicles.

As the future directions, the system's performance can be further enhanced by porting the trained networks on more advanced and powerful edge devices thus tailoring it for real-time onboard deployments. Moreover, the current system focuses on object recognition, but it can be further trained and modified to incorporate image segmentation, road and lane detection, traffic signal and road signs classification, and object tracking for providing comprehensive driver assistance.

Acknowledgment

The authors would like to acknowledge Cosmin Rotariu from Xperi-Ireland and the rest of the team members for providing the support in preparing the data accusation setup and helping throughout in data collection and Quentin Noir from Lynred France for giving their feedback. Moreover, the authors would like to acknowledge the contributors of all the public datasets for providing the image resources to carry out this research work and ultralytics for sharing the YOLO-v5 Pytorch version.

References

[1] Heliaus European Union Project, Accessed: Jan. 15, 2022. [Online]. Available: https://www.heliaus.eu/

[2] Nvidia Jetson Nano, Accessed: Jan. 18, 2022. [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano-developer-kit

[3] Nvidia Jetson Xavier NX Development kit, Accessed: Jan. 18, 2022. [Online]. Available: https://developer.nvidia.com/embedded/jetson-xavier-nx-devkit

[4] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *Proc. 2005 7th IEEE Workshops Appl. Comput. Vis. (WACV/MOTION'05)*, vol. 1, 2005, pp. 364–369, doi: 10.1109/ACVMOT.2005.14.

[5] CVC-09 FIR Sequence Pedestrian Dataset, Accessed: Jan. 20, 2022. [Online]. Available: http://adas.cvc.uab.es/elektra/enigma-portfolio/item-1/

[6] A. Torabi, G. Massé, and G. A. Bilodeau, "An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications," *Comput. Vis. Image Understanding*, vol. 116, no. 2, pp. 210–211, 2012, doi: 10.1016/j.cviu.2011.10.006.

[7] Z. Wu, N. Fuller, D. Theriault, and M. Betke, "A thermal infrared video benchmark for visual analysis," *IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 201–208, 2014, doi: 10.1109/CVPRW.2014.39.

[8] Z. Xu, J. Zhuang, Q. Liu, J. Zhou, and S. Peng, "Benchmarking a large-scale FIR dataset for on-road pedestrian detection," *Infrared Phys. Technol.*, vol. 96, pp. 199–208, 2019, doi: 10.1016/j.infrared.2018.11.007.

[9] FLIR Thermal Dataset, Accessed: Jan. 29, 2022. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/

[10] Y. Choi *et al.*, "KAIST multi-spectral day/night data set for autonomous and assisted driving," in *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 934–948, Mar. 2018, doi: 10.1109/TITS.2018.2791533.

[11] Q. Liu, Z. He, X. Li, and Y. Zheng, "PTB-TIR: A thermal infrared pedestrian tracking benchmark," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 666–675, Mar. 2020, doi: 10.1109/TMM.2019.2932615.

[12] R. D. Brehar, M. P. Muresan, T. Mariţa, C. -C. Vancea, M. Negru, and S. Nedevschi, "Pedestrian street-cross action recognition in monocular far infrared sequences," *IEEE Access*, vol. 9, pp. 74 302–74 324, 2021, doi: 10.1109/ACCESS.2021.3080822.

[13] R. R. Ziyatdinov and R. A. Biktimirov, "Automated system of recognition of road signs for ADAS systems," in *Proc. IOP Conf. Ser., Mater. Sci. Eng.*, vol. 412, no. 1, p. 012081, 2018, doi: 10.1088/1757-899X/412/1/012081.

[14] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The german traffic sign detection benchmark," in *Proc. Int. Joint Conf. Neural Netw.*, 2013, pp. 1–8, doi: 10.1109/IJCNN.2013.6706807.

[15] O. Töro, T. Bécsi, and S. Aradi, "Performance evaluation of a bernoulli filter based multi-vehicle cooperative object detection," *Transp. Res. Procedia*, vol. 27, pp. 77–84 2017, doi: 10.1016/j.trpro.2017.12.042.

[16] G. Yan, M. Yu, Y. Yu, and L. Fan, "Real-time vehicle detection using histograms of oriented gradients and adaboost classification," *Optik (Stuttg)*, vol. 127, no. 19, pp. 7941–7951, 2016, doi: 10.1016/j.ijleo.2016.05.092.

[17] A. M. Ali, W. I. Eltarhouni, and K. A. Bozed, "On-road vehicle detection using support vector machine and decision tree classifications," in *Proc. 6th Int. Conf. Eng. MIS*, 2020, pp. 1–5, doi: 10.1145/3410352.3410803.

[18] L. Zhang, J. Tan, D. Han, and H. Zhu, "From machine learning to deep learning: Progress in machine intelligence for rational drug discovery," *Drug Discov. Today*, vol. 22, no. 11, pp. 1680–1685, 2017, doi: 10.1016/j.drudis.2017.08.010.

[19] Pretrained object detection and classification models, Accessed: Jan. 29, 2022. [Online]. Available: https://pytorch.org/serve/model_zoo.html

[20] A. Corovic, V. Ilic, S. Duric, M. Marijan, and B. Pavkovic, "The real-time detection of traffic participants using YOLO algorithm," *26th Telecommun. Forum*, 2018, pp. 1–4, doi: 10.1109/TELFOR.2018.8611986.

[21] V. Ghenescu, E. Barnoviciu, S. V. Carata, M. Ghenescu, R. Mihaescu, and M. Chindea, "Object recognition on long range thermal image using state of the art DNN," *Conf. Grid, High Perform. Comput. Sci.*, 2018, pp. 1–4, doi: 10.1109/ROLCG.2018.8572026.

[22] A. Narayanan, R. Darshan Kumar, R. Roselinkiruba, and T. Sree Sharmila, "Study and analysis of pedestrian detection in thermal images using YOLO and SVM," *6th Int. Conf. Wireless Commun., Sig. Process. Netw.*, pp. 431–434, 2021, doi: 10.1109/WiSPNET51692.2021.9419443.

[23] R. Kalita, A. K. Talukdar, and K. K. Sarma, "Real-Time human detection with thermal camera feed using YOLOv3," *IEEE 17th India Council Int. Conf.*, 2020, pp. 1–5, doi: 10.1109/INDICON49873.2020.9342089.

[24] M. Ivašić-Kos, M. Krišto, and M. Pobar, "Human detection in thermal imaging using YOLO," in *Proc. ACM Int. Conf. Proc. Ser.*, vol. Part F148262, 2019, pp. 20–24, doi: 10.1145/3323933.3324076.

[25] M. P. Muresan, S. Nedevschi, and R. Danescu, "Robust data association using fusion of data-driven and engineered features for real-time pedestrian tracking in thermal images," *Sensors*, vol. 21, no. 23, p. 8005, 2021, doi: 10.3390/s21238005.

[26] X. Han, J. Chang, and K. Wang, "Real-time object detection based on YOLO-v2 for tiny vehicle object," *Procedia Comput. Sci.*, vol. 183, pp. 61–72, 2021, doi: 10.1016/j.procs.2021.02.031.

[27] V. Paidi, H. Fleyeh, and R. G. Nyberg, "Deep learning-based vehicle occupancy detection in an open parking IoT using thermal camera," *IET Intell. Transp. Syst.*, vol. 14, no. 10, pp. 1295–1302, 2020, doi: 10.1049/iet-its.2019.0468.

[28] M. Kristo, M. Ivasic-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using YOLO," *IEEE Access*, vol. 8, pp. 125 459–125 476, 2020, doi: 10.1109/ACCESS.2020.3007481.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[30] Yolo version camparison, Accessed: Feb. 1, 2022. [Online]. Available: https://pjreddie.com/darknet/yolo

[31] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020. [Online]. Available: arxiv.org/abs/2004.10934

[32] YOLO-v5 link, Accessed: Feb. 1, 2022. [Online]. Available: https://github.com/ultralytics/yolov5

[33] M. A. Farooq, P. Corcoran, C. Rotariu, and W. Shariff, "Object detection in thermal spectrum for advanced driver-assistance systems (ADAS)," *IEEE Access*, vol. 9, pp. 156 465–156 481, 2021, doi: 10.1109/ACCESS.2021.3129150.

[34] A. Farouk Khalifa, E. Badr, and H. N. Elmahdy, "A survey on human detection surveillance systems for raspberry pi," *Image Vis. Comput.*, vol. 85, pp. 1–13, 2019, doi: 10.1016/j.imavis.2019.02.010.

[35] D. S. Breland, S. B. Skriubakken, A. Dayal, A. Jha, P. K. Yalavarthy, and L. R. Cenkeramaddi, "Deep learning-based sign language digits recognition from thermal images with edge computing system," *IEEE Sens. J.*, vol. 21, no. 9, pp. 10 445–10 453, May 2021, doi: 10.1109/JSEN.2021.3061608.

[36] Y. Liu, S. Cao, P. Lasang, and S. Shen, "Modular lightweight network for road object detection using a feature fusion approach," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 51, no. 8, pp. 4716–4728, Aug. 2021, doi: 10.1109/TSMC.2019.2945053.

[37] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361, doi: 10.1109/CVPR.2012.6248074.

[38] G. Balasekaran, S. Jayakumar, and R. Pérez de Prado, "An intelligent task scheduling mechanism for autonomous vehicles via deep learning," *Energies*, vol. 14, no. 6, p. 1788, 2021, doi: 10.3390/en14061788.

[39] A. Saragaglia and A. Durand, "Method of infrared image processing for non-uniformity correction." U.S. Patent No. 10,015,425, Jul. 2018.

[40] Lynred France, Accessed: Feb. 5, 2022. [Online]. Available: https://www.lynred.com/

[41] J. de Vries, "Image processing and noise reduction techniques for thermographic images from large-scale industrial fires," *Quantitative InfraRed Thermogr.*, Bordeaux, France, Jul. 2014.

[42] LabelImg object annotation tool, Accessed: Jan. 15, 2022. [Online]. Available: https://github.com/tzutalin/labelImg

[43] M. A. Ganaie and M. Hu, "Ensemble deep learning: A review," 2021. [Online]. Available: https://arxiv.org/abs/2104.02395

[44] Jetson statistics tool, Accessed: Jan. 31, 2022. [Online]. Available: https://github.com/rbonghi/jetson_stats

[45] Nvidia TensorRT for developers, Accessed: Feb. 6, 2022. [Online]. Available: https://developer.nvidia.com/tensorrt

Muhammad Ali Farooq received the B.E. degree in electronic engineering from IQRA University, Karachi, Pakistan, in 2012, and the M.S. degree in electrical control engineering from the National University of Sciences and Technology, Islamabad, Pakistan, in 2017. He is currently working toward the Ph.D. degree with the National University of Ireland Galway (NUIG), Galway, Ireland. His research interests include machine vision, computer vision, video analytics, and sensor fusion. He has won the prestigious H2020 European Union (EU) scholarship and is currently working with NUIG as one of the consortium partners in the Heliaus (thermal vision augmented awareness) project funded by EU.

Waseem Shariff received the B.E. degree in computer science from the Nagarjuna College of Engineering and Technology, Bangalore, India, in 2019, and the M.S. degree in computer science, specializing in artificial intelligence from the National University of Ireland Galway (NUIG), Galway, Ireland, in 2020. He is currently with NUIG. He is Associated with Heliaus (thermal vision augmented awareness) Project. He is also allied with FotoNation/Xperi research team. His research interests include machine learning utilizing deep neural networks for computer vision applications, including working with synthetic data, thermal data, and RGB.

Peter Corcoran (Fellow, IEEE) is currently the Personal Chair of electronic engineering with the College of Science and Engineering, National University of Ireland Galway, Galway, Ireland. He was the Co-Founder in several start-up companies, notably FotoNation, now the Imaging Division of Xperi Corporation. He has more than 600 cited technical publications, more than 120 peer-reviewed journal articles, 160 international conference papers, and a coinventor on more than 300 granted U.S. patents. He is an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He is a Member of the IEEE Consumer Technology Society for more than 25 years and the Founding Editor of the *IEEE Consumer Electronics Magazine*.

Appendix G

Infrared Imaging for Human Thermography and Breast Tumor Classification using Thermal Images

Authors' Contribution to [20]

Contribution Criteria	*Contribution Percentage*
Research Hypothesis	MAF: 100%
Experiments and Implementation	MAF: 100%
Background	MAF: 100%
Manuscript Preparation	MAF: 80%, PC: 20%

Infrared Imaging for Human Thermography and Breast Tumor Classification using Thermal Images

Muhammad Ali Farooq
College of Engineering and Informatics
National University of Ireland Galway (NUIG)
Galway, Ireland
m.farooq3@nuigalway.ie

Peter Corcoran
College of Engineering and Informatics
National University of Ireland Galway (NUIG)
Galway, Ireland
peter.corcoran@nuigalway.ie

***Abstract*— Human thermography is considered to be an integral medical diagnostic tool for detecting heat patterns and measuring quantitative temperature data of the human body. It can be used in conjunction with other medical diagnostic procedures for getting comprehensive medication results. In the proposed study we have highlighted the significance of Infrared Thermography (IRT) and the role of machine learning in thermal medical image analysis for human health monitoring and various disease diagnosis in preliminary stages. The first part of the proposed study provides comprehensive information about the application of IRT in the diagnosis of various diseases such as skin and breast cancer detection in preliminary stages, dry eye syndromes, and ocular issues, liver disease, diabetes diagnosis and last but not least the novel COVID-19 virus. Whereas in the second phase we have proposed an autonomous breast tumor classification system using thermal breast images by employing state of the art Convolution Neural Network (CNN). The system achieves the overall accuracy of 80% and recall rate of 83.33%.**

Keywords—Infrared Thermography, Deep Neural Networks, Thermal camera, Computer Aided Dignosis, Classification

I. Introduction

Thermal imaging is one of the most rapidly growing imaging techniques nowadays [1]. It can be described as a key method for measuring the spatial temperature of various materials, objects, and scenes. It plays a pivotal role in detecting abnormal temperature patterns of the human body. It works by absorbing IR radiations emitted from the human body and then generating heat energy indications with or without visible illumination conditions. The heat maps are generated in different color schemes such as iron, grayscale and rainbow thermals maps. These color maps are generally used to define different temperature ranges which eventually help us in identifying the health parameters of the human body. Fig. 1 shows the heat map in five different color maps of the thermal human face of the healthy subject. These images are acquired using an uncooled prototype thermal camera developed under the Heliaus EU project [29].

Thermography is the most common technique for acquiring valuable information using thermal cameras [2]. It collects information using an array of infrared sensors to read infrared energy emissions (surface temperature) to determine the operating conditions of different parts of the human body. It consists of two main components: Thermo and Graphy where Thermo refers to temperature patterns of the body and Graphy refer to image acquisition techniques as shown in Fig. 2.

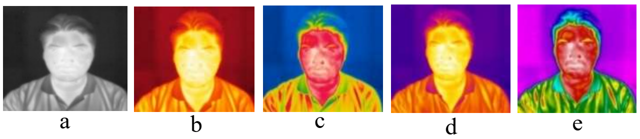


Fig. 1. Visualization of different temperature color maps of a thermal facial image a) greyscale, b) glow, c) HSV, d) iron, e) rainbow.

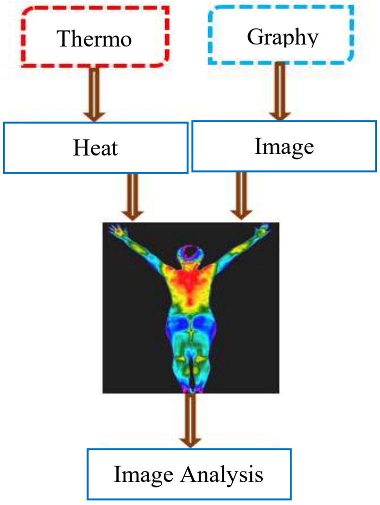


Fig. 2. Main components of human thermography.

Thermographic cameras usually detect radiation in the long-infrared range of the electromagnetic spectrum (roughly 9,000– 14,000 nm or 9–14 μm) and produce images using that radiation, which is generally referred to as thermograms. The amount of radiation emitted by an object increases with an increase in the temperature; therefore, thermography helps in analyzing the minute temperature variation patterns. In the overall classification, thermography can be divided into two main types which include active thermography and passive thermography. The passive thermography works by pointing the IR camera at the investigated body and checking whether the investigated body is at a lower or higher temperature than the background. Whereas, the active thermography approach is based on the excitation of the sample by applying external energy into it and subsequently measuring the thermal response from it. Therefore, active thermography is a fully dynamic process requiring different methods of image processing [3].

II. Background Research

The human body temperature is considered as a vital parameter that can be used for human health monitoring and various disease diagnosis. Measuring temperature with thermal camera systems comes with advantages such as non-contact and non-invasive diagnostic procedures. Moreover, thermography is a patient-friendly method and does not only provide conventional temperature measurement but also gives a comprehensive image of a patient's body temperature distribution which can be ultimately used to extract essential information regarding the overall body health [2]. Human

978-1-7281-9418-9/20/$31.00 ©2020 IEEE

clinical thermography depends on the exact examination of the skin surface temperature as an impression of the typical or anomalous physiology of the human [4]. There is a wide range of conventional medical imaging methods that give the field of internal imaging of the human body from outside inspection. For instance, we can remotely infuse and follow the radioactive isotopes to the body. The process of radiology works by utilizing the x-ray beam integrating with matter for plotting the internal structure of the body, yet these methods have weaknesses like low affectability and risk of harmful x-ray radiations, particularly when being emanated for a long time [5]. In comparison to this, infrared thermography is a patient-friendly technique, that provides a non-invasive detailed temperature distribution of human body. Thus it does not have destructive impacts of the harmful radiation which can be caused by conventional medical imaging procedures like x-ray, gamma rays, and Computed Tomography (CT) scans. In short, thermal imaging benefits us my analyzing abnormal temperature patterns of the human body that are the natural indication of any type of disease [6, 7]. However, the use of infrared thermal imaging in humans is dependent on different factors that need to be considered while acquiring such datasets. It includes environmental factors, individual factors dependent on human body intrinsic and extrinsic characteristics, and last but not least technical factors such as camera calibration, field of view (FoV) and subject distance from camera [20].

III. Human Thermography for Various disease Diagnosis

This section will mainly focus on various human disease detection using human thermography based health monitoring system and diagnostic tools.

A. Infrared Thermography for Cancer Detection

Cancer or carcinoma tumors can be defined as one of the most fatal diseases in the human body. It is generally due to the abnormal growth of cells in any specific part of the body. It has the possibility of spreading to other parts of the body which can eventually lead to more serious medical conditions thus making it untreatable. Therefore, the detection of cancer in preliminary stages is a prime objective that allows doctors and specialists to perform specialized medical treatments to eventually cure the patients. Conventional medical procedures such as biopsy tests to check blood samples of infected areas of the body are often very painful for the patients. However, thermography can be efficiently used to detect different types of cancer in any part of the body. The process is painless as it provides non-contact and non-invasive diagnostic procedures. The process of thermography simply works by detecting the higher temperature in specific parts of the body thus the radiation compression also increases. It is due to more amount of heat generated from abnormal cancerous cells. Tumors can cause an increment in metabolism rate and blood flow which transports local stains with high temperatures in place that can be easily detected via the process of infrared thermography [2, 8]. IRT can be effectively used for different types of cancer detection in early stages which includes breast cancer detection [2], skin cancer detection [9], and brain tumor diagnosis [10].

B. Infrared Thermography for Diabetes Diagnosis

Diabetes is the most rapidly growing disease in middle and low-income countries [27]. The more severe stages can cause paralysis and leg issues. The main reasons for these issues are low blood flow referred to as vascular disorder and the loss of feeling or weakness also knowns as neuropathy in medical terminology. During such type of disease patients normally undergoes abnormal skin temperature, thus making thermography an appropriate tool for diagnosis of vascular disorder or neuropathology. Such types of abnormal thermal patterns happen in the patient's leg and hands like temperature decrement in foot and toes. Generally, a diabetic patient suffers from higher temperature with average thermal readings of about 30.2 ± 1.3°C [15]. Therefore, infrared thermography plays a vital role in the initial diagnosis of diabetes in the human body which will eventually aid the doctors and specialists to provide appropriate treatment to their patients [11].

Vision Quest [12, 30] has developed a thermal optical imaging system capable of detecting early symptoms of diabetic peripheral neuropathy in the plantar foot, which accounts for about 25% of hospital stays among diabetes patients. The system works by recording the thermalized video of the patient's foot during recovery from cold provocation. The overall system comprises of high end and low noise infrared camera which is periodically calibrated to minimize thermal sensitivity to less than 0.50 °C. The system works by extracting the post images referred to as functional signals to detect dynamic changes in microvascular blood flow which are then analyzed. According to their initial research, the system can show visible statistical and significant differences between normal patients and subjects who have been diagnosed with peripheral neuropathy.

C. Thermography for Diagnosis of Liver Disease

Thermal Imaging especially near-infrared imaging is widely used for prodromal detection of chronic liver diseases. One such type of disease is liver fibrosis. It is a pathological process that can escalate to a more severe stage medically referred to as cirrhosis which eventually results in liver failure at its final stages. It is considered to be one of the major public health concerns that affect hundreds of millions of people in both developed and developing countries [13]. Therefore, early detection of liver fibrosis is of prime cause thus preventing them from the development of cirrhosis with chronic liver disease. Conventionally the level of fibrosis is examined by histological assessment using Mason's Trichrome stain performed by two different senior pathologists in a single-blind test and the severity of fibrosis is measured using Meavir Score shown in Table I.

TABLE I. Fibrosis Severity Scale

S:No	Level	Description
1.	F0	No fibrosis
2.	F1	Mild fibrosis
3.	F2	Moderate fibrosis
4.	F3-F4	Advanced fibrosis

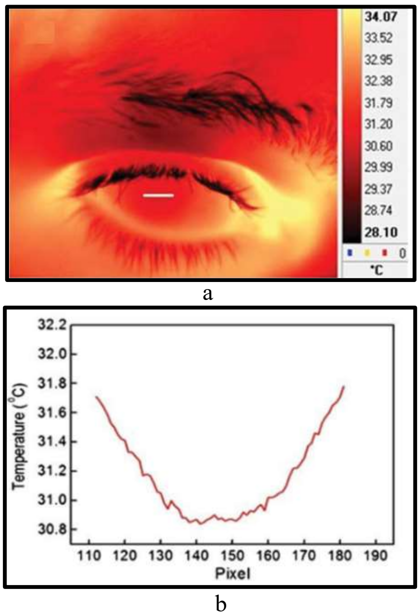
But the process is time-consuming and requires years of experience for correct diagnosis. In fibrosis, De novo formation of such blood vessels can increase the surface temperature of the human liver however if the fibrosis

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826131. The JU receives support from the European Union's Horizon 2020 research and innovation program and National funding from France, Germany, Ireland (Enterprise Ireland International Research Fund), and Italy.

advances to further stages i.e F3 cirrhosis, an excessive accumulation of connective tissue is observed in the liver and it results in the decrement of the surface temperature of the organ. These abnormal patterns of thermal temperature in lever can be easily detected by thermal imaging cameras which can be eventually used for the early diagnosis of liver fibrosis with high precision thus curing the patients from growing it into advanced stages [13].

D. *Thermography for Eye Ocular Issues*

Thermology is used in the field of human ophthalmology for the diagnosis of dry eye syndromes and ocular issues [14] by observing the eye physiology. The process of non-intrusive Infrared thermography (IRT) works by detecting the abnormal temperature behaviors of dry eye which is nearly about (32.38 ± 0.69°C). It is slightly higher as compared to the temperature of a healthy eye which is about (31.94 ± 0.54 °C) [15]. Generally, the horizontal temperature distribution in heathy eye organ is symmetrical and it is relatively low in the geometric center of the cornea as shown in Fig. 3 (image reproduced by the author's permission).



a

b

Fig. 3. Thermographic test image of a human eye, a) Horizontal temperature distribution of healthy cornea, b) Horizontal line in the graph indicates the thermal characteristics [15].

E. *Infrared Imaging for Detecting Novel COVID-19 Virus*

Currently, coronavirus has become one of the largest widespread disease throughout the world. According to the reported statistics and figures [16] around 194 countries [16] have been affected with more than 5.1 million cases all over the world and more than 330,000 [16] people died due to it. Medically it has been termed as COVID-19 and it has been declared as a pandemic from the World Health Organization (WHO) [28]. The symptoms of this disease appear in 2-14 days and the immune system of the affected person detects an infection that results in raise of core body temperature. Other symptoms of this virus include dry cough, tiredness and last but not least shortness of breath. Since high temperature is one of the prime symptoms [31] of this disease thermal imaging devices can be effectively used to detect the elevated temperature pattern in the human body. Numerous airports around the world [17] have installed thermal imaging cameras also referred to as heat scanners for the robust screening of the passengers. Further image processing and computer vision based algorithms are used to generate a color palette that represents different temperature scales that aids in the diagnosis of this virus. In the wake of widespread of COVID-19 virus, FLIR [18] is experiencing increased demand for its hand-held T-series products as well as its A310 fixed-mounted thermal imaging camera [19].

IV. Role of Image Processing and Machine Learning for Effective Human Thermography

Thermal waves are exponentially reduced in an environment, and hence the thermal effects of abnormalities are often subtle. Moreover, thermal images also suffer from a relatively low signal-to-noise ratio (SNR) [2]. Thus, digital image processing plays a vital role in providing reliable solution such that by applying a variety of filters in both frequency and time domain to overcome these factors [2, 32]. Digital image processing techniques are used offline to enhance the quality of low quality pre-recorded thermal images of the human body to better visualize the image from both human and machine perspective. It works by providing dynamic contrast control, edge preservation [2], and removing unwanted noise from the image by applying different algorithms such as applying various filtering methods, thresholding techniques and, probabilistic models. Once the images are refined, the enhnaced outputs of thermal imaginary datasets can be fed into a variety of machine learning algorithms to extract meaningful information which can ultimately help us to detect any type of abnormalities in the human body. Fig. 4 illustrates the generic comprehensive block diagram representation of a thermal imaging based Computer Aided Dignosis (CAD) system.

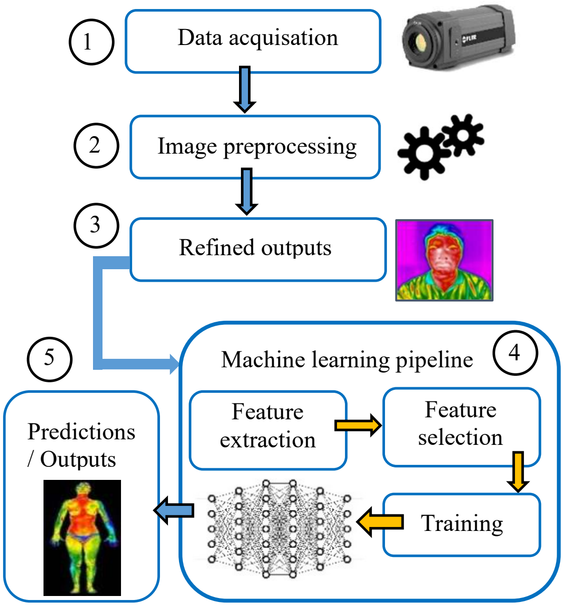


Fig. 4. Comprehensive block diagram representation of thermal imaging based Computer Aided Dignosis (CAD) system.

As illustrated in Fig. 4 the overall system works by acquiring images using different types of thermal cameras such as mobile thermal cameras, LWIR thermal cameras, and NIR thermal cameras. In the next step, the acquired data is processed using image processing algorithms to produce refined outputs. Refined outputs are then fed as input data in a variety of machine learning algorithms to extract important feature values. Conventional machine learning classifiers such as Support Vector Machines (SVM) and Naive Bayes mainly rely on handcrafted features that are engineered manually using a variety of feature extractors thus having chances of higher error rate. Also, the SVM classifier is not computationally efficient when dealing with large datasets. As a solution to this Deep Neural Networks (DNN) plays a prime role since it uses learned feature values that are self extracted from raw pixel images. Therefore, DNN benefits us by providing high accuracy and mitigating the drawbacks handcrafted feature engineering. DNN is extensively used for classification and segmentation applications in medical imaging to provide the second opinion to doctors and specialists. To train these networks, different types of network hyperparameters are used to achieve optimal generalization and regularization in DNN networks. It includes the selection of appropriate error function, optimizers, learning rate, momentum, batch size, and the number of iterations. Finally, the networks are trained to achieve precise and robust accuracy levels which are validated by performing cross-validation on unseen test data.

V. Proposed Methodology for Breast Cancer Classification using Deep Learning

In this section, we have proposed a breast tumor classification system using thermography images of breast cancer by applying deep learning methodologies as discussed in Section IV. It is the most common type of cancer throughout the world and found very commonly in women. In 2020 it is expected that about 276,480 new cases of invasive breast cancer are to be diagnosed in women only in the U.S along with 48,540 new cases of non-invasive (in situ) breast cancer [26]. However, if it is detected in preliminary stages it is treatable by taking suitable medical measures.

In our study, we have utilized DMR - Database for Mastology Research [21]. It is a type of online platform that stores and manages mastologic images for early detection of breast cancer. The dataset is consisting of different modalities of breast cancer images which include thermography images, mammography images, MRI images, and ultrasound images. The dataset was collected using FLIR SC-620 camera [22] from 287 patients of different age groups. The overall dataset includes images using static and dynamic data acquisition protocols. In static data acquisition set up the body of the patient must achieve thermal balance in a controlled environment whereas dynamic protocols are used to inspect the skin temperature recovery caused by thermal stress after cooling the patient by electric fan. For the proposed study we have used data acquired through the dynamic protocol as it provides extensive thermal data as compared to static data. Dynamic data acquisition provides a set of 20 images and 2 additional lateral images of each patient which was acquired during a certain interval of time. Considering the dynamic methodology, we have used data of 40 patients, among which 18 patients belong to the cancerous class and 22 patients belong to the healthy (benign) set. We have utilized a pre-trained Inception-v3 deep neural network [23] network for effective classification between benign and cancerous cases. Fig. 5 shows the complete workflow diagram of the proposed system.

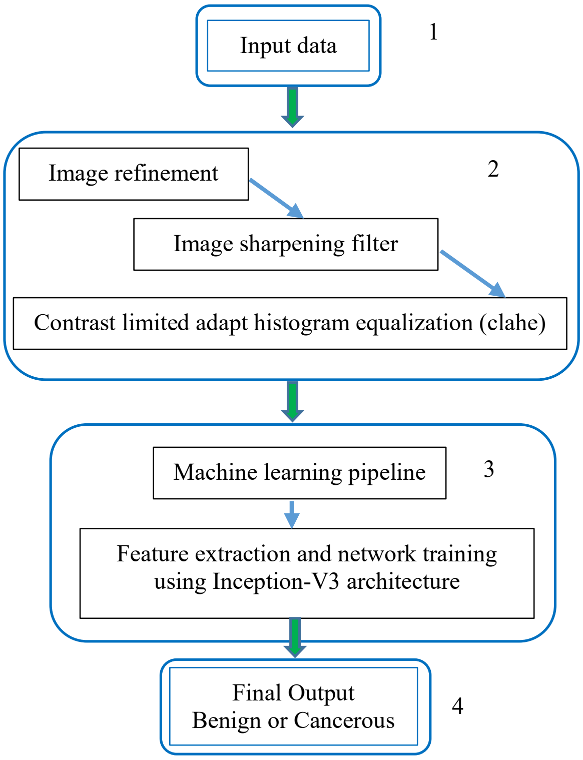

Input data
1
Image refinement
2
Image sharpening filter
Contrast limited adapt histogram equalization (clahe)
Machine learning pipeline
3
Feature extraction and network training using Inception-V3 architecture
Final Output
Benign or Cancerous
4

Fig. 5. Complete workflow diagram for autonomous breast cancer classification system.

A. Image Processing

As shown in Fig. 5, in the first phase system works by performing the image preprocessing operations which includes applying initial sharpening filter and then applying Contrast Limited Adapt Histogram Equalization (CLAHE) operations on original images provided in the dataset. The sharpening filter is used to increase the contrast in the images, especially where different color channels meet. CLAHE operation is used for performing intensity normalization in the image. It works by taking different parameters which include distribution and clip limit. Distribution specifies the spreading scale that histogram equalization will utilize as the basis for generating the contrast transform function. Clip limit is generally defined as an overall contrast enhancement threshold limit. We have used uniform distribution function that creates a flat histogram and clip limit is set to 0.01. The main purpose of applying the image preprocessing operation is to refine the existing image quality by making the high-level features more descriptive which will be further used for training the CNN network. The same preprocessing techniques are applied to whole training data. Fig. 6 shows the preprocessing operations applied to one of the test cases from DMR - Database for Mastology Research [21].

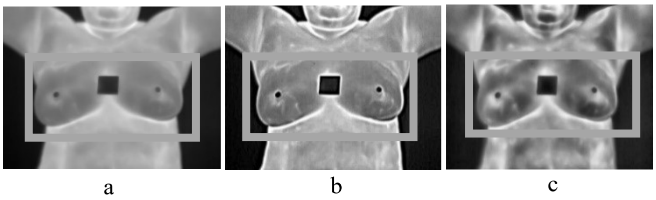


Fig. 6. Preprocessing operations applied on thermographic breast tumor image a) original image, b) sharpened image, c) contrast limited adapt histogram equalization image.

B. *Deep Neural Network Training*

In the second stage, the processed images along with original images are used for training the state-of-the-art inception-v3 [23] network. It is a 48-layer deep neural network designed by Google Brain which was initially trained on ImageNet library [24]. The main reason for employing this network, it uses multiple features from multiple filters which improve the overall performance of the network. Moreover, all the established architectures before the inception network performed convolution on the spatial and channel-wise domain together. By performing the 1x1 convolution, the inception block is doing cross-channel correlations, thus ignoring the spatial dimensions. It is then followed by cross-spatial and cross-channel correlations using the 3x3 and 5x5 filters. In the proposed study we have used the weights of the pre-trained inception -v3 network and retrained the last layers of the network for our custom breast tumor classification task by applying transfer learning.

VI. Experimental Results

The overall algorithm is implemented using Core I7 sixth-generation machine equipped with NVIDIA RTX 2080 Graphical Processing Unit (GPU) having 8GB of dedicated graphic memory. As discussed in Section V we have first applied image preprocessing operations to refine the original images provided in the dataset. The processed images along with original images (input image size of 299 x 299) are used for training the state-of-the-art inception-v3 [23] network using TensorFlow deep learning platform as exhibited in Fig. 7.

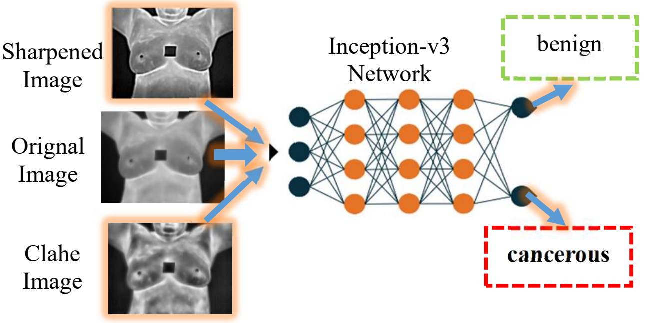


Fig. 7. Training of inception-v3 architecture for breast tumor classification.

The data is distributed in the ratio of 70%, 20% and 10% for training, validation, and testing purposes in an empirical fashion. After getting unsatisfactory training and validation experimental results in the initial stages, the following set of network hyperparameters is selected as shown in Table II to avoid model overfitting and achieve optimal generalization.

TABLE II. Inception-v3 Training Parameters

Epochs	Learning Rate	Batch size	Optimizer	Error function
5000	0.001	32	Stochastic Gradient Decent (SGD)	Binary Cross-Entropy

The system achieves the overall training accuracy of 93.73% and validation accuracy of 91.32%. Fig. 8 shows the accuracy and loss graph of the inception-v3 network.

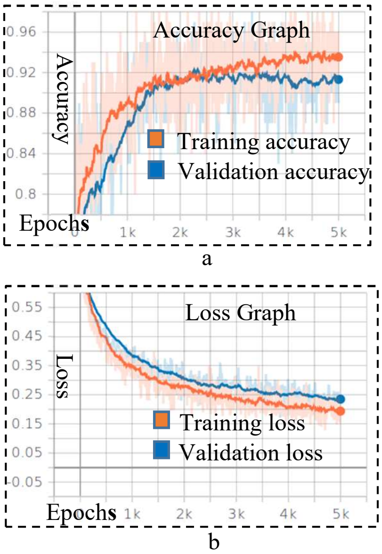
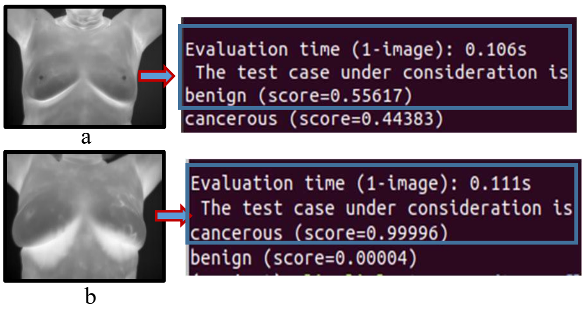


Fig. 8. Inception-v3 network training and loss graphs a) training and validation accuracy graph, b) loss graph of inception-v3 network.

The classifier is then cross-validated on unseen test cases to check the overall test accuracy of the network. It is important to mention that the trained network is tested without applying any of the preprocessing operations on test data to validate the overall robustness of inception-v3 architecture. Fig. 9 shows the results of the correct prediction on two random test cases along with the confidence scores and individual inference time required. The overall performance of the inception-v3 network on test data has been evaluated using five different quantitative measures which include accuracy, sensitivity, specificity, precision, and F1 score [25]. The results of these metrics are shown in Table III.

TABLE III. Quantitively Metrics Results

Metrics		Score
1.	Accuracy	80%
2.	Sensitivity /Recall	83.33%
3.	Specificity	77.77%
4.	Precision	71.43%
5.	F1 Score	76.89%



```
Evaluation time (1-image): 0.106s
 The test case under consideration is
benign (score=0.55617)
cancerous (score=0.44383)
```

```
Evaluation time (1-image): 0.111s
 The test case under consideration is
cancerous (score=0.99996)
benign (score=0.00004)
```

Fig. 9. Inference results on two random correct cases a) Patient 1 prediction: benign case with a confidence score of 55.61% and inference time 0.106 second, b) Patient 2 prediction: cancerous case with a confidence score of 99.99 % and inference time of 0.111 second.

VII. Conclusion and Future Work

The main objective of the entire study is to emphasize the importance of thermography and role of machine learning in thermal medical image analysis for human health monitoring and disease diagnosis in prodromal stages. Technologically advanced platforms for performing effective human thermography are also referred to as Computer-Aided Diagnosis System (CAD) and is considered to be a sixth sense and reliable second opinion for doctors, specialists and medical experts. This has been evident in our study by proposing a breast tumor classification system using grayscale thermal images. The system works by employing sate of art inception-v3 architecture for performing precise classification between benign and malignant (cancerous) cases. The system achieves the overall accuracy of 80% and the sensitivity of 83.33%.

For future prospects, we believe that extensive use of smartphone-based and commercial grade thermal cameras could make human thermographic data widely accessible for investigating in depth details in this area. Moreover, advanced computational techniques such as machine learning and deep learning algorithms can be integrated with existing thermal cameras hardware to come up with the concept of smart thermal diagnosis systems. Such types of systems can be deployed in cars for in-cabin Driver Monitoring Systems (DMS) for making correct predictions about the driver's health promptly.

ACKNOWLEDGMENT

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826131. The JU receives support from the European Union’s Horizon 2020 research and innovation program and National funding from France, Germany, Ireland (Enterprise Ireland International Research Fund), and Italy.

REFERENCES

[1] Fitzgerald, Anita, and Jessica Berentson-Shaw. "Thermography as a screening and diagnostic tool: a systematic review." NZ Med J 125.1351 (2012): 80-91.

[2] Lahiri BB, S Bagavathiappan, T Jayakumar, John Philip. “Medical applications of infrared thermography: a review”. Infrared Physics & Technology 2012;55:221-35.

[3] Wiecek, B. "Review on thermal image processing for passive and active thermography." *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2006.

[4] Ju Xiangyang, Jean Christophe Nebel, J Paul Siebert. “3D thermography imaging standardization technique for inflammation diagnosis”. In Infrared Components and Their Applications 2005;5640:266-74.

[5] Kim, Dong Ik. "PREFACE: How Dangerous Are X-ray Studies That We Undertake Every Day?." Journal of Korean medical science 31.Suppl 1 (2016): S2-S3.

[6] Hardy James D, Carl Muschenheim. Radiation of heat from the human body. V. The transmission of infra-red radiation through skin. J Clin Invest 1936;15:1-9.

[7] Yang WJ, Yang PP. Literature survey on biomedical applications of thermography. Biomed Mater Eng 1992;2:7-18.

[8] Kennedy Deborah A, Tanya Lee, Dugald Seely. A comparative review of thermography as a breast cancer screening technique. Integr Cancer Ther 2009;8:9-16

[9] Narayanamurthy, Vigneswaran, et al. "Skin cancer detection using non-invasive techniques." RSC advances 8.49 (2018): 28095-28130.

[10] Gorbach, Alexander M., et al. "Intraoperative infrared imaging of brain tumors." Journal of neurosurgery 101.6 (2004): 960-969.

[11] Selvarani, A., and G. R. Suresh. "Infrared Thermal Imaging for Diabetes Detection and Measurement." Journal of medical systems 43.2 (2019): 23.

[12] P Soliz, C Agurto, A Edwards, Z Jarry, J Simon, M Burge, “Detection of diabetic peripheral neuropathy using spatial-temporal analysis in infrared videos,” Asilomar, Nov 2016

[13] M. G. Ramírez-Elías, E. S. Kolosovas-Machuca, D. Kershenobich, C. Guzmán, G. Escobedo, and F. J. González, “Evaluation of liver fibrosis using Raman spectroscopy and infrared thermography: A pilot study,” *Photodiagnosis Photodyn. Ther.*, vol. 19, no. September, pp. 278–283, 2017.

[14] Tan Jen Hong, E YK Ng, U Rajendra Acharya, Caroline Chee. Infrared thermography on ocular surface temperature: a review. Infrared physics & technology 2009;52:97-108.

[15] H. Shirzadfar, F. Ghasemi, and M. Shahbazi, “A Review of Recent Application of Medical Thermography in Human Body for Medical Diagnosis,” vol. 2, no. 2, pp. 102–120, 2018.

[16] Covid 19 Current Stastistics, Weblink: https://www.worldometers.info/coronavirus/, (last accesssed on March 26, 2020).

[17] Covid-19 Virus, Passenger Screening on Airports, Weblink: https://www.airport-technology.com/news/coronavirus-airports-screening-passengers-china/, (last accssed on March 26, 2020)

[18] FLIR Thermal Homepage Weblink: https://www.flir.eu/, (Last accessed on March 23, 2020).

[19] Increase in demand of FLIR Thermal Camera due to Global Corona Virus, Weblink:https://www.photonics.com/Articles/Demand_for_FLIR_Temperature_Screening_Devices/a65632, (Last accessed on March 25, 2020)

[20] Fernández-Cuevas, Ismael, et al. "Classification of factors influencing the use of infrared thermography in humans: A review." Infrared Physics & Technology 71 (2015): 28-55.

[21] Silva, L. F., et al. "A new database for breast research with infrared image." Journal of Medical Imaging and Health Informatics 4.1 (2014): 92-100.

[22] FLIR SC-620 thermal camera datasheet, Weblink: http://www.gammadata.se/assets/Uploads/SC620-specsheet.pdf, (Last accessed on March 26, 2020).

[23] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[24] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[25] M. Stojanovi et.al., “Understanding sensitivity, specificity, and predictive values”, Vojnosanit Pregl, vol. 71, no11, pp. 1062–1065,2014.

[26] Breast Cancer Statistics Weblink:https://www.breastcancer.org/symptoms/understand_bc/statistics?gclid=CjwKCAiAlO7uBRANEiwA_vXQ-0MaLVshK8MzLKyRZZadEmrXMsWuAxzT4vVrCTD-ACXefsmT4PEcSxoCemkQAvD_BwE, (Last accessed on March 15, 2020).

[27] World Health Organization Diabetes facts and figures, Weblink: https://www.who.int/news-room/fact-sheets/detail/diabetes, (Last accessed on March 26, 2020).

[28] Covid-19 Global Pandemic, Weblink: https://www.who.int/emergencies/diseases/novel-coronavirus-2019, (Last accessed on March 29, 2020).

[29] Heliaus European Union Project Website: https://www.heliaus.eu/, (Last accessed on 20th January 2020).

[30] Vision Quest Biomedical System, Weblink: http://visionquest-bio.com/diabetic-neuropathy.aspx, (Last accessed on May 8, 2020).

[31] Corona Virus Symptoms, Weblink: https://www.nhs.uk/conditions/coronavirus-covid-19/check-if-you-have-coronavirus-symptoms/, (Last accessed on May 8, 2020).

[32] Saha, Arindam, et al. "ThermoFlowScan: Automatic Thermal Flow Analysis of Machines from Infrared Video." *VISIGRAPP (4: VISAPP)*. 2017.

[33] Canziani, Alfredo, Adam Paszke, and Eugenio Culurciello. "An analysis of deep neural network models for practical applications." arXiv preprint arXiv:1605.07678 (2016).

Appendix H

Advanced Deep Learning Methodologies for Skin Cancer Classification in Prodromal Stages

Authors' Contribution to [21]

Contribution Criteria	*Contribution Percentage*
Research Hypothesis	MAF: 80%, PC: 20%
Experiments and Implementation	MAF: 100%
Background	MAF: 100%
Manuscript Preparation	MAF: 70%, AK: 10%, VV: 10% PC: 10%

Advanced Deep Learning Methodologies for Skin Cancer Classification in Prodromal Stages

Muhammad Ali Farooq[1], Asma Khatoon[1], Viktor Varkarakis[1], Peter Corcoran[1]

[1] National University of Ireland (NUIG) Galway H91CF50, IRELAND
{m.farooq3, a.khatoon1, v.varkarakis1, peter.corcoran}
@nuigalway.ie

Abstract. Technology-assisted platforms provide reliable solutions in almost every field these days. One such important application in the medical field is the skin cancer classification in preliminary stages that need sensitive and precise data analysis. For the proposed study the Kaggle skin cancer dataset is utilized. The proposed study consists of two main phases. In the first phase, the images are preprocessed to remove the clutters thus producing a refined version of training images. To achieve that, a sharpening filter is applied followed by a hair removal algorithm. Different image quality measurement metrics including Peak Signal to Noise (PSNR), Mean Square Error (MSE), Maximum Absolute Squared Deviation (MXERR) and Energy Ratio/ Ratio of Squared Norms (L2RAT) are used to compare the overall image quality before and after applying preprocessing operations. The results from the aforementioned image quality metrics prove that image quality is not compromised however it is upgraded by applying the preprocessing operations. The second phase of the proposed research work incorporates deep learning methodologies that play an imperative role in accurate, precise and robust classification of the lesion mole. This has been reflected by using two state of the art deep learning models: Inception-v3 and MobileNet. The experimental results demonstrate notable improvement in train and validation accuracy by using the refined version of images of both the networks, however, the Inception-v3 network was able to achieve better validation accuracy thus it was finally selected to evaluate it on test data. The final test accuracy using state of art Inception-v3 network was 86%.

Keywords: Melanoma, CNN, DNN, Dermoscopy, Inception-v3, MobileNet,

1 Introduction

Cancer nowadays is one of the greatest growing groups of diseases throughout the world, among which skin cancer is most common of them. According to stats and figures, the annual rate of skin cancer is increasing at an alarming rate each year [1]. The modern medical science and treatment procedures prove that if skin cancer is detected in its initial phase then it is treatable by using appropriate medical measures which includes laser surgery or removing that part of the skin which ultimately could save a patient's life. Skin cancer has two main stages which include malignancy and melanoma among which melanoma is fatal and comes with the highest risk. In most cases, malignant mole is clearly visible on the patient's skin which is often identified by the patients themselves.

Dermoscopic diagnosis refers to a non-invasive skin imaging method, which has become a core tool in the diagnosis of melanoma and other pigmented skin lesions. However, performing dermoscopy using conventional methods may lower down the diagnostic accuracy which can lead to more chances of errors. These errors are generally caused by the complexity of lesion structures and the subjectivity of visual interpretations [18].

Computer-Aided Diagnosis (CAD) system is a type of digitized platform based on advanced computer vision, deep learning, and pattern recognition techniques for skin cancer classification. For the proposed study we have designed a CAD system for skin cancer classification by utilizing advanced deep neural networks. The system consists of the following steps: Firstly, a preprocessing of the digital images which includes removing clutter such as hair from that part of the skin where the pigmented mole is present and applying a sharpening filter to make that area more clear and visible thus minimizing the chances of error. The next essential step includes the feature extraction and classification process to extract the results for the cases under consideration by utilizing deep learning techniques. Section 2 presents the background and related study and highlights the medical aspects regarding skin cancer. Section 3 describes the detailed methodology of the proposed system whereas Section 4 presents the implementation and experimental results of the proposed study. Section 5 draws the overall conclusion of the paper.

2 Background/ Related Work

The human skin is the largest organ of the overall human body. It covers all other organs of the body. It guards the entire body from microbes, bacterium, ultraviolet radiation, helps to regulate body temperature and permits the sensations of touch, heat, and cold [2].

2.1 Skin Moles and Skin Cancer

Mole or nevus on human skin can be described as a dark, erected spot comprised of skin cells that are grown in a group rather than individually. These cells are generally known as melanocytes which are responsible for producing melanin, the pigment color in our skin. The main reason behind mole development on human skin is predominantly because of direct sun exposure and any kind of extreme injury. The fair skin population has a greater ratio of skin moles due to the lower quantity of melanin (natural pigments) in their skins [3]. There are three different kinds of skin malignant growth, which include Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), and Melanoma. Malignancy is a description of the "stage" of cancer. These malignant growths are critical however, Melanoma comes with the highest risk level and it is discovered more frequently in individuals maturing under 50 years for men and over 50 years for women [4].

2.2 Related Work/ Previous Studies

The study proposed by Simon Kalouche utilizes [5] computer vision-based deep learning methods to detect skin cancer and more specifically melanoma. Their dataset was trained on 3 different learning models including a logistic regression model, fine-tuned VGG-16 and multi-layer perceptron deep neural network to achieve a significant amount of classification accuracy. Their results show that their algorithm ability to segment moles and classify skin lesions is 70% to 78%. Md Zahangir et al [6] presented the Inception Recurrent Residual Convolutional Neural Network (IRRCNN) method for breast cancer classification on BreakHis and some other publicly available datasets. They compared their experimental results against existing machine learning techniques in terms of patch-based, image-based, patient-level and image-level classification. Their IRRCNN models provide efficient classification in terms of Area Under the Curve (AUC), global accuracy and the ROC curve. Andre Esteva et al [7] demonstrated classification of skin cancer using a single CNN, which they have achieved by end to end training using image data which is based on disease and pixel labels. In their work, they utilized a large dataset of clinical images that consist of several diseases.

3 Methodology

In the proposed study, an efficient skin cancer diagnosis system has been implemented for precise classification between malignant melanoma and benign cases. The complete algorithm consists of several steps starting from the input phase of applying image preprocessing ranging to the analysis of the case under consideration in the form of the probability of lesion Malignancy. Fig. 1 shows the complete workflow of the proposed algorithm.

3.1 Image Preprocessing

For the proposed study, the Kaggle skin cancer dataset [8] consisting of processed skin cancer images of ISIC Archive [9] has been utilized. The dataset has a total of 2637 training images and 660 testing images with a resolution of 224 x 224. It is consists of two main classes which include melanoma and benign cases. For image preprocessing two major operations have been applied which includes an initial sharpening filter followed by hair removal filter using dull razor software [10]. These were selected in order to remove the clutter. The results of the image preprocessing operations on two random sample cases are shown in Fig. 2.

It is noteworthy that image quality is refined after applying the image preprocessing operations. This is shown in Section 4 of the paper where the results from four different image quality metrics Peak Signal to Noise Ratio (PSNR), Mean Squared Error (MSE), Maximum Absolute Squared Deviation (MXERR) (MXERR) and Ratio of Squared Norms (L2RAT) on both ground truth images, and preprocessed images are presented.

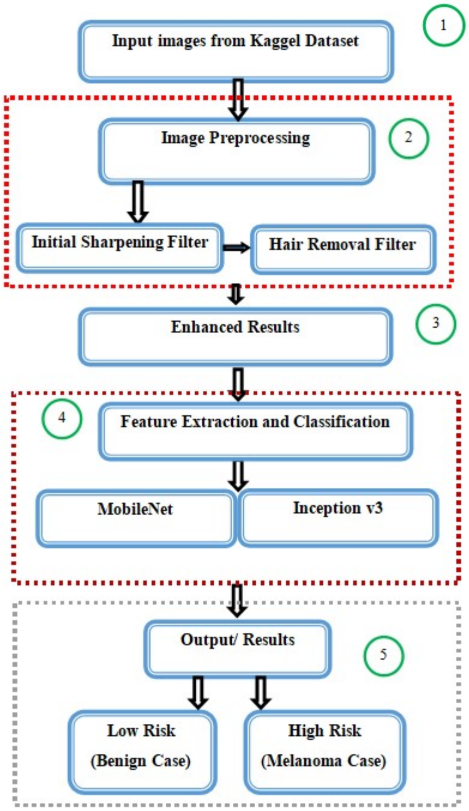


Fig 1. Workflow diagram of the proposed method

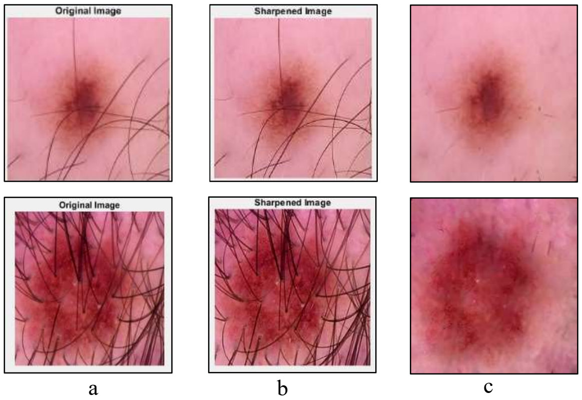


Fig. 2. Image preprocessing operations a) original image, b) initial Matlab sharpening filter, c) hair removal using dull razor software [10]

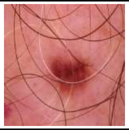
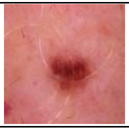
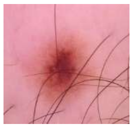
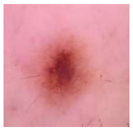
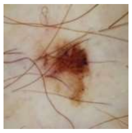
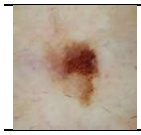
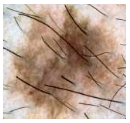
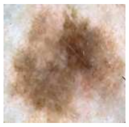
3.2 Feature Extraction and Classification

In the next step, the processed images are fed to state-of-the-art deep neural networks in order to perform the feature extraction and classification steps. In this work, the Inception-v3 and MobileNet deep learning architectures are utilized. These architectures play a vital role by extracting feature values from raw pixel images. The Inception-v3 has state of the art performance in the classification task. It is made up of 48 layers stacked on top of each other [11]. The Inception-v3 model was initially trained using 1.2 million images from Imagenet [12] of 1000 different categories. These pre-trained layers have a strong generalization power and they are able to find and summarize information that will help to classify most of the images from the real-world environment. For the proposed study we have utilized this network for our custom classification task by retraining the final layer of the network thus updating and finetuning the softmax layer, by applying the method of transfer learning. This was preferred as the amount of data available for this task is limited and training the Inception-v3 from the beginning would require a lot of time and computational resources. Therefore, by fine-tuning the inception v3 model, we take advantage of its powerful pre-trained layers and thus being able to provide satisfying accuracy results even with a limited amount of data. MobileNet is one of the other finest deep learning architectures proposed by Howard et al. 2017 [13] specifically designed for mobile and embedded vision applications. MobileNet is counted as a lightweight deep learning architecture. It uses depth-wise separable convolutions that means it performs a single convolution on each color channel rather than combining all three and flattening it. This has the effect of filtering the input channels. For our experiments, the networks were trained with two different types of data. The networks were trained with the original images and also with the images after applying the preprocessing operations to them. The training and validation accuracy were examined in order to study the effect of the training on the networks with the two different types of data. Finally, the accuracy on the test set is calculated in order to evaluate the overall performance of the classifiers

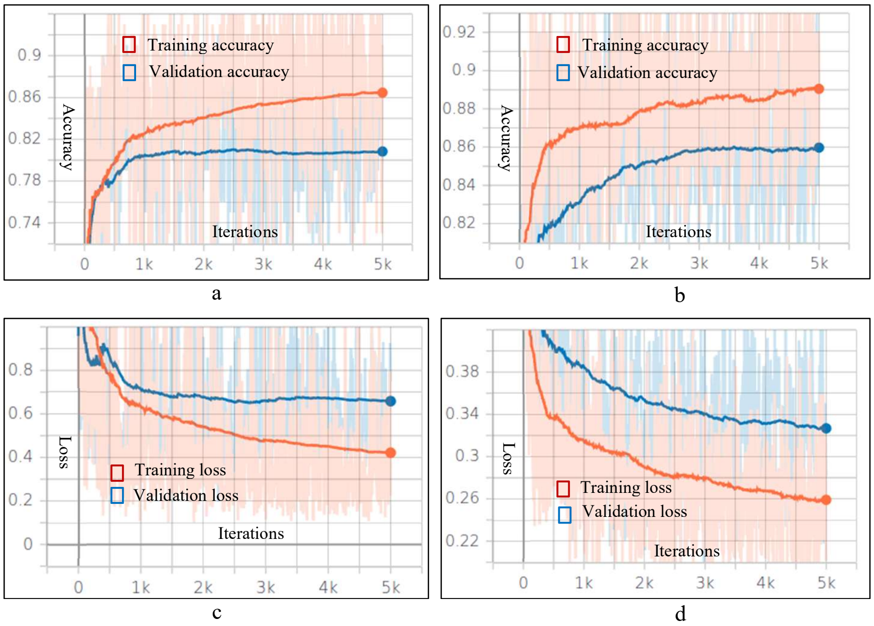
4 Implementation and Experimental Result

The overall algorithm was implemented using Matlab R2018a for computing image quality metrics and TensorFlow [14] for training the classifiers. The system was trained and tested on a Core I7 sixth-generation machine equipped with NVIDIA RTX 2080 Graphical Processing Unit (GPU) having 8GB of dedicated graphic memory. The first part of the experimental results displays the image quality metrics measured for both benign and malignant melanoma cases before and after applying the image preprocessing operations. It is displayed in Table 1.

Table 1. Image Quality Metrics

Image	PSNR	MSE	MAXERR	L2RAT	Dimension
	19.4205	743.0656	99	0.9657	224 x 224
	21.5481	655.2738	99	0.9801	224 x 224
	22.1285	398.3229	99	0.9868	224 x 224
	23.2953	304.4737	99	0.9902	224 x 224
	22.4291	371.6785	99	0.9852	224 x 224
	24.0840	329.9128	99	0.9903	224 x 224
	18.6732	882.5930	99	0.9681	224 x 224
	19.3975	847.0221	99	0.9747	224 x 224

The experimental results show clearly that image quality is not comprised however it is upgraded which is evident from high PSNR values and other metrics after applying image preprocessing operations especially the hair removal filter. The image quality metrics were carried out on more than fifty images and the same observations were measured. The second part of the experiments includes the training of the classifiers using the two state of the art deep learning networks i.e. Inception-v3 and MobileNet. For Inception-V3 the data was resized to 299 x 299 since the network has an image input size of 299 by 299. The classifiers were trained on both sets of images i.e. original (ground truth) images and images after applying the preprocessing operations to them. Both the networks were trained using the same hyperparameters. The learning rate was set to 0.005 with a batch size of 32 and total iterations were set to 5000. The training data was split in the ratio of 75% and 25% for training and validations images respectively. Fig. 3 and Fig. 4 display the training and validation accuracy graphs along with the error rate (cross-entropy) graph of MobileNet and Inception-v3 networks.



a

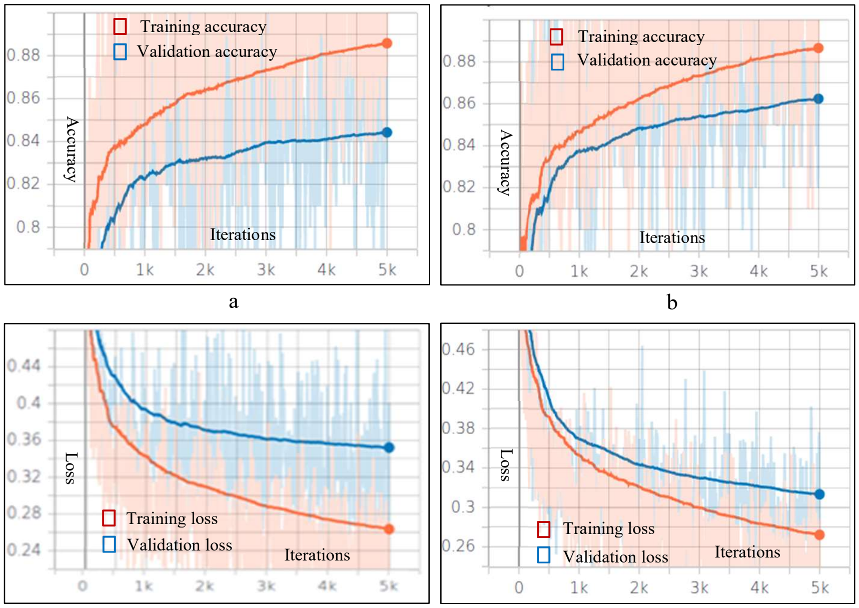
b

c

d

Fig. 3. Accuracy and loss graph of MobileNet network a), training and validation accuracy before applying image preprocessing operations b), training and validation accuracy after applying image preprocessing operations c), training and validation loss before applying image preprocessing operations and d) training and validation loss after applying image preprocessing operations

The accuracy graphs in Fig. 3 show that training and validation accuracy before applying the image preprocessing was 86% and 79.8% and it was increased to 89% and 85.9% by using a refined version of images obtained after applying the image preprocessing operations. Similarly, the validation error rate was also decreased from 61% to 32% by using the refined version of images.



a

b

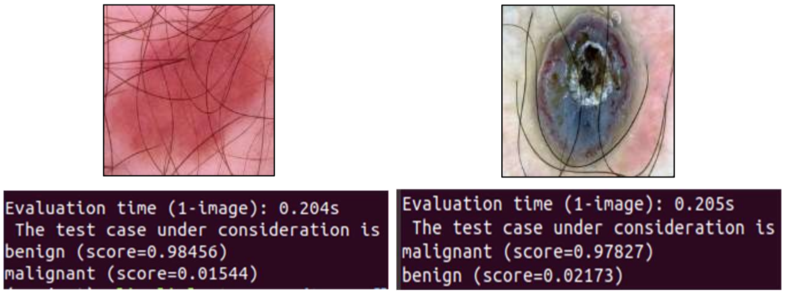
c

d

Fig. 4. Accuracy and loss graph of Inception-v3 network a), training and validation accuracy before applying image preprocessing operations b), training and validation accuracy after applying image preprocessing operations c), training and validation loss before applying image preprocessing operations and d) training and validation loss after applying image preprocessing operations

The accuracy graphs in Fig. 4 show that training and validation accuracy before applying the image preprocessing was 88.3% and 84.2%. By using the refined version of images training accuracy tends to remain the same thought the validation accuracy was increased to 86.1%. Similarly, the validation error rate was also decreased from 36% to 32.3% by using the refined version of images.

Overall, in both networks, significant improvements were measured after using the refined version of images. The experimental results show that the Inception-v3 network was able to achieve better validation accuracy using a refined version of training data i.e. 86.1 % thus we will be using the Inception-v3 network for evaluating it on the test data. For evaluating the classifiers on the test data, we have picked numerous cases from the test set from both classes, benign and malignant melanoma among which visually complex and challenging test cases were selected for the proposed research work. It is pertinent to mention that the network was tested using the original images (unrefined version) to test the overall effectiveness of the classifier. Fig. 5 shows some of the results predicted correctly on test images. Table 2 illustrates the complete results on visually complex test cases selected for the proposed study which will be further used to evaluate overall testing accuracy, sensitivity (true positive rate), specificity (true negative rate) and precision metrics. The rows highlighted with red color indicates the misclassified test cases when compared with ground truth results.



```
Evaluation time (1-image): 0.204s
 The test case under consideration is
benign (score=0.98456)
malignant (score=0.01544)
```

a

```
Evaluation time (1-image): 0.205s
 The test case under consideration is
malignant (score=0.97827)
benign (score=0.02173)
```

b

Fig. 5. Test case results on two random cases using Inception-v3 network a) case 4 – (benign = Low risk = 98.4% confidence level), b) Case 16 – (malignant melanoma = high risk = 97.8% confidence level).

Table 2. Individual Test Case Results

Test Case	Predicted results using Inception-v3 Network trained on original images	Predicted results using Inception-v3 Network trained on processed images	Ground truth Results
1	Benign – Low risk – 97.8%	Benign – Low risk – 84.6%	Low risk
2	Malignant–High risk – 91.8%	Malignant – High risk – 89.1%	High risk
3	Benign – Low risk – 98.4%	Benign – Low risk – 96.9%	Low risk
4	Benign – Low risk – 98.4%	Benign – Low risk – 98.4%	Low risk
5	Malignant – High risk – 96.4%	Malignant – High risk – 95.7%	Low risk
6	Malignant – High risk – 98.7%	Malignant – High risk – 96.2%	High risk
7	Malignant – High risk – 98.8%	Malignant – High risk – 97.8%	High risk
8	Malignant – High risk – 99.4%	Malignant – High risk – 99.3%	High risk
9	Benign – Low risk – 71.2%	Benign – Low risk – 60.7%	Low risk
10	Malignant – High risk – 85.9%	Malignant – High risk – 76.8%	Low risk
11	Malignant – High risk – 99.5%	Malignant – High risk – 99.3%	High risk
12	Malignant – High risk – 98.5%	Malignant – High risk – 99.2%	High risk
13	Malignant – High risk – 70.9%	Malignant – High risk – 87.4%	High risk
14	Benign– Low risk – 74.8%	Malignant – High risk – 92.3 %	High risk

15	Malignant – High risk – 97.5 %	Malignant – High risk – 98.7 %	High risk
16	Malignant – High risk – 99.0 %	Malignant – High risk – 97.8 %	High risk
17	Benign – Low risk – 96.1%	Benign – Low risk – 97.0 %	Low risk
18	Benign – Low risk – 86.4%	Benign – Low risk – 86.2 %	Low risk
19	Benign – Low risk – 99.4%	Benign – Low risk – 99.1 %	Low risk
20	Malignant –High Risk– 50.2%	Benign – Low risk – 59.0 %	Low risk
21	Malignant – High risk – 81.9 %	Benign – Low risk – 64.1 %	High risk

The overall performance of the Inception-v3 network on test data has been evaluated using five quantitative measures: Accuracy, sensitivity, specificity, precision and F1 score [15,19]. These measures are computed using the following forms.

$$Accuracy\ (ACC) = \frac{tp + tn}{tp + tn + fp + fn}\ X\ 100 \qquad (1)$$

$$Sensitivity\ (TPR)/\ Recall = \frac{tp}{tp + fn}\ X\ 100 \qquad (2)$$

$$Specificity\ (TNR) = \frac{tn}{tn + fp}\ X\ 100 \qquad (3)$$

$$Precesion\ (PPV) = \frac{tp}{tp + fp}\ X\ 100 \qquad (4)$$

$$F1\ Score\ = 2\ X\ \frac{Precesion\ X\ Recall}{Precesion + Recall}\ X\ 100 \qquad (5)$$

Where t_p,f_p, f_n, and t_n refer to true positive, false positive, false negative, and true negative. ACC in (1) means overall testing accuracy, TPR in (2) means true positive rate, TNR in (3) refers to true negative rate while PPV in (4) is an abbreviation for positive prediction value.

Table 3 illustrates the results of all the four quantitative measures: Accuracy, sensitivity, specificity, and precision of the Inception-v3 network before and after using the image preprocessing operations on test data. It can be observed that testing accuracy is increased to 86% by training the classifier using the refined version of images.

Table 3. Overall Quantitative Metrics Results on Test Data

Quantitative measures	Inception-v3 network trained on original images	Inception-v3 Network trained on refined (processed) images
Accuracy	81%	86%
Sensitivity	87.5%	89%
Specificity	77%	83%
Precision	70%	80%
F1 Score	77%	84%

5. Conclusion and Future work

The main purpose of the proposed study was to improve the overall accuracy level of two state of art deep learning networks which include Inception-v3 and MobileNet by using the refined version of skin cancer images obtained after applying image preprocessing operations. The experiments were conducted using the Kaggle Skin Cancer Dataset by applying initial sharpening filter and hair removal algorithms. Initially, we applied these algorithms as image pre-processing mechanisms to remove the clutters thus producing the refined version of images. Different image quality metrics including Peak Signal to Noise (PSNR), Mean Square Error (MSE), Maximum Absolute Squared Deviation (MXERR) and Energy Ratio/ Ratio of Squared Norms (L2RAT) were used to compare the image quality before and after applying the pre-processing techniques. These metrics prove that image quality was upgraded after applying sharpening filter and hair removal algorithms. In the next phase of experimental results, we have seen substantial improvement in training, validation and test accuracy after applying image pre-processing operation. Thus, we have achieved an overall test accuracy of 86% using state of the art Inception-v3 network by fine-tuning the last layer of the network with a refined version of kaggle skin cancer training dataset.

For future work, more image pre-processing techniques like neural networks based super image algorithms and other such techniques could be used to improve the image quality to a better extent. Moreover, other state of the art deep neural networks such as ResNet-101 [16], Xception [17] could be utilized in order to improve the accuracy levels.

References

1. Skin Cancer Facts and Figures, Web Link: http://www.skincancer.org/skin-cancer-information/skin-cancer-facts, (Last accessed on 04th Oct 2019).
2. Organs of the Body, Web Link: http://hubpages.com/education/Human-Skin-The-largest-organ-of-the-Integumentary-System, (Last accessed on 05th Oct 2019).
3. Types of Skin Moles, Web Link: https://skinvision.com/en/articles/types-of-skin-moles-and-how-to-know-if-they-re-safe, (Last accessed on 01st Oct 2019)

4. Skin Cancer Risk Factors. Web Link: http://www.cancer.org/cancer/skincancer-melanoma/detailedguide/melanoma-skin-cancer-risk-factors, (Last Accessed on 01st Oct 2019).
5. Kalouche, Simon, Andrew Ng, and John Duchi. "Vision-based classification of skin cancer using deep learning." 2015, conducted on Stanfords Machine Learning course (CS 229) taught (2016).
6. Alom, Md Zahangir, et al. "Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network." Journal of digital imaging (2019): 1-13.
7. Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." Nature 542.7639 (2017): 115.
8. Kaggle Skin Cancer Dataset, Web Link: https://www.kaggle.com/fanconic/skin-cancer-malignant-vs-benign, (Last accessed on 24th September 2019).
9. ISIC Archive Dataset,Web Link: https:// www.isic archive.com, (Last accessed on 24th September 2019).
10. Lee T, Ng V, Gallagher R, Coldman A, McLean D. DullRazor: “A software approach to hair removal from images” Published in the Computers in Biology and Medicine, 1997
11. Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
12. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012
13. Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).
14. TensorFlow Deep Learning Platform, Web Site: https:// www.tensorflow.org, (Last accessed on 27th September 2019).
15. M. Stojanovi et.al., “Understanding sensitivity, specificity, and predictive values”, Vojnosanit Pregl, vol. 71, no11, pp. 1062–1065,2014.
16. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
17. Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
18. Doi, Kunio. "Computer-aided diagnosis in medical imaging: historical review, current status, and future potential." Published in *Computerized medical imaging and graphics 31.4 (2007): 198-211*, 2007.
19. F1 Score in Machine Learning, Web Link: https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1, (Last accessed on 29th Oct 2019).

Appendix I

Towards Monocular Neural Facial Depth Estimation: Past, Present, and Future

Authors' Contribution to [22]

Contribution Criteria	*Contribution Percentage*
Research Hypothesis	FK: 70%, MAF: 20%, PC: 10%
Experiments and Implementation	FK: 100%
Background	FK: 70%, WS: 30%
Manuscript Preparation	FK: 60%, MAF: 15%, WS: 10%, PC: 10%, SB: 5%

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 28, 2022, accepted March 9, 2022, date of publication March 11, 2022, date of current version March 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158950

Towards Monocular Neural Facial Depth Estimation: Past, Present, and Future

FAISAL KHAN[1], MUHAMMAD ALI FAROOQ[1], WASEEM SHARIFF[1], SHUBHAJIT BASAK[2], (Graduate Student Member, IEEE), AND PETER CORCORAN[1], (Fellow, IEEE)

[1]Department of Electronic Engineering, College of Science and Engineering, National University of Ireland Galway, Galway, H91 TK33 Ireland
[2]School of Computer Science, National University of Ireland Galway, Galway, H91 TK33 Ireland

Corresponding author: Faisal Khan (f.khan4@nuigalway.ie)

This work was supported in part by the College of Science and Engineering, National University of Ireland Galway, Galway, Ireland; in part by Xperi Galway, Galway; and in part by the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant 18/CRT/6224.

ABSTRACT This article contains all of the information needed to conduct a study on monocular facial depth estimation problems. A brief literature review and applications on facial depth map research were offered first, followed by a comprehensive evaluation of publicly available facial depth datasets and widely used loss functions. The key properties and characteristics of each facial depth map dataset are described and evaluated. Furthermore, facial depth maps loss functions are briefly discussed, which will make it easier to train neural facial depth models on a variety of datasets for both short- and long-range depth maps. The network's design and components are essential, but its effectiveness is largely determined by how it is trained, which necessitates a large dataset and a suitable loss function. Implementation details of how neural depth networks work and their corresponding evaluation matrices are presented and explained. In addition, an SoA neural model for facial depth estimation is proposed, along with a detailed comparison evaluation and, where feasible, direct comparison of facial depth estimation methods to serve as a foundation for a proposed model that is utilized. The model employed shows better performance compared with current state-of-the-art methods when tested across four datasets. The new loss function used in the proposed method helps the network to learn the facial regions resulting in an accurate depth prediction. The network is trained on synthetic human facial depth datasets whereas for validation purposes real as well as synthetic facial images are used. The results prove that the trained network outperforms current state-of-the-art networks performances, thus setting up a new baseline method for facial depth estimations.

INDEX TERMS Facial depth datasets, loss functions, neural depth estimation, empirical and systematic evaluation.

I. INTRODUCTION

The process of obtaining 3D information from a 2D frame is known as depth estimation. Depth estimation is used in diversified computer vision applications such as augmented reality, posture estimation, 3D reconstruction, object detection and recognition, semantic segmentation and -human-machine interaction, weather forecast, and autonomous vehicles. The ground truth depth information used to estimate depth is beneficial for developing reliable navigation systems for intelligent vehicles, environmental reconstruction, and image interpretation to understand the objects in the image and the scene behind them.

Face depth estimation is a challenging subject that has been explored in conjunction with face motion [1], facial analysis, and facial recognition [2], [3]. Many methods for estimating face depth have been presented in recent years, notably 3D from stereo replicating [4], 3D morphable model-based methods [5], [6], shape from shading (SfS) [5], [6], shape from motion techniques (SfM) [6], [7], and statistical techniques [8], [9]. Due to the facial symmetry of facial areas, the stereo matching procedure for face depth estimation is more complicated (regardless of utilizing the local or global technique), particularly when the system is binocular and therefore only one stereo pair is used. Stereo matching

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li.

methods can estimate a reasonable depth or disparity map for facial depth estimation, but these approaches are more sophisticated, requiring the use of a local or global procedure. Because of the similarity of the face areas, particularly when using a binocular setup with only one pair of stereo images. All stereo approaches are limited by the similarity characteristics of the facial information. Furthermore, the similarity of the pixels values results in more spikes, holes, and particularly uncertain disparities in the depth map.

The computer vision field has conventionally approached the field of depth maps in a variety of methods, such as with stereo or multi-view cameras [10], [11], structure from motion [12], [13], and depth from light diffusion & shading [14], [15]. The described methods face many difficulties, such as missing pixel values and depth consistency, which result in inconsistencies in depth maps. In addition, the camera calibration, camera setup, and post-processing techniques are computationally expensive and time-consuming. The research community has explored the monocular depth estimation task using only a single image which is much more straightforward and suitable for consumer applications. The credit goes to significant advances in machine learning-based networks [16]–[20]. In the first part of the paper, we have given a detailed evaluation of publicly available facial depth datasets and widely used loss functions in facial depth estimation networks, thus to better understanding the problem of facial depth maps. The key characteristics and properties of the facial depth datasets are presented and compared, followed by the loss functions employed. The implementation specifics of how neural depth networks work, as well as the evaluation matrices that correlate to them, are shown and described. A full comparison evaluation and, where possible, direct comparison of facial depth estimation methods are performed in the second phase of the paper to serve as a foundation for a proposed model that is used. When tested across four datasets, the proposed model outperforms current state-of-the-art approaches. The suggested method's unique loss function aids the network in learning the facial areas, resulting in an accurate depth prediction. The network is trained using synthetic human facial depth datasets, and real and synthetic facial images from four facial depth datasets are used for validation.

A. RESEARCH CONTRIBUTIONS

Following thorough research over the previous few years, image-based facial depth estimation using deep learning algorithms has demonstrated promising results. However, the field is still in its early stages, and more improvements are expected to address issues and challenges such as data selection for training, generalization to unknown environments, fine-scale depth estimation, reconstruction versus recognition, handling multiple objects in the presence of occlusions, and cluttered backgrounds, data imbalance and how to select an appropriate loss function and neural model for facial depth estimation.

This paper aims to provide all of the key information for conducting a study on monocular facial depth estimation challenges. First, a brief review of the literature and applications of facial depth map research was presented, followed by a detailed analysis of publicly available facial depth datasets and commonly used loss functions. To better understand the facial depth map problem, the facial depth dataset's key characteristics and properties are described and evaluated, followed by the loss functions used. For each dataset, the dataset description, metadata, ground truth, and relevant data (year of publishing, ground truth information, image size, type, objects per image, and several images) are listed systematically. In addition, each loss function is presented in such a way that the research community can select the best loss function for their requirements. The implementation details of how neural depth networks work are demonstrated and explained, as are the evaluation matrices that correspond to them. In the second section of the paper, a complete comparison evaluation and, where possible, direct comparison of facial depth estimation methods are conducted to serve as a foundation for a proposed model that is used. The model outperforms current state-of-the-art techniques when tested across four datasets. The unique loss function of the suggested method supports the network in learning the facial areas, resulting in an accurate depth prediction. The network is trained with synthetic human facial depth datasets and validated with real and synthetic facial images from four facial depth datasets.

B. CHALLENGES AND DEVELOPMENTS

Monocular facial depth estimation based on deep learning (DL) has been intensively explored and advanced over the last few years. However, still, several limitations need to be addressed. This section covers the major issues and discusses potential directions for monocular facial depth estimation maps research. By utilizing a deep learning network, we can extract many features simultaneously, such as semantic information, optical flow features, and depth features. While semantic segmentation will be incorporated into depth estimation, it will remain a separate module that performs autonomous tasks. Additionally, there are typically numerous sub-networks capable of learning depth estimation, visual odometry, and flow estimation. However, such networks are not adequately connected, which results in a large set of network parameters, which eventually requires an increased memory footprint. How to improve the network's integration is a research direction that is worth exploring as the future direction of this research work.

The quality of the training data has a significant impact on the generalization and reliability of the deep learning model. To increase facial depth estimation accuracy, more data with higher quality and a wider variety of scene types is required. However, the facial depth estimation datasets currently available are quite small, and creating a new dataset is time-intensive and expensive. At the moment, several researchers generate a large number of images for facial depth estimation using a variety of software, but the quality is inconsistent. A future research goal will be to provide a dataset for monocular facial depth estimation that is compatible with deep learning models.

Realistic environments are frequently complex, having a high amount of moving objects, occlusions, changing light conditions, and changing weather. However, the majority of existing facial depth estimation models assume an optimum environment. Although some researchers have attempted to address dynamic objects and occlusion scenarios and have made considerable progress lately, the problem of improving the facial depth estimation of complicated scenes for real-world applications remains a key future research field.

Facial depth estimation is a challenging stage in the development of practical applications such as augmented reality (AR), virtual reality (VR), robotics and autonomous vehicles. However, the resolution of the estimated facial depth is often limited in most existing facial depth estimation algorithms to maximize computational effectiveness.

The fundamental module of SLAM is image depth estimation, which is deeply connected with commercial applications such as autonomous driving. However, researchers frequently design deeper networks with more parameters and constraints to accomplish depth estimation, which needs more computational cost and hence does not fulfil the real-time requirements of modern applications. Thus, a future research area will be to determine how to use a lighter network for real-time estimation while maintaining prediction accuracy.

The rest of the paper is organized as follows: Section 2 discusses related work in the domain of facial depth estimation, especially related studies, or surveys. Section 3 presents the results of a bibliometric investigation, a thorough examination of depth datasets, and further discusses the most used loss functions. Section 4 presents the implementation details of how facial depth neural networks work followed by some comparative analysis of the facial depth estimation methods. Section 5 presents evaluation matrices and section 6 describes and illustrates the most recent SoA depth estimation model, which is discussed and chosen for facial depth estimation. Section 7 shows the experimental results, discusses the training approach, and compares the trained model to SoA methods in a brief comparison study. Section 8 includes a detailed discussion of the experimental results while section 9 provides the conclusion and future research directions.

II. RELATED WORKS

Datasets are the foundations for evaluating the behaviour and validating the results of artificial intelligence networks, and they play a critical role in scientific research. Another important building block is to use an appropriate loss function to improve the deep network's training performance. An in-depth analysis of various facial depth datasets is performed, and depth regression loss functions for both short and long-range depth datasets are proposed in the next sections. This section focuses mostly on related facial depth estimation research and applications.

A. FACIAL DEPTH ESTIMATION APPLICATIONS

Human face images are among the most common images, and they play an important role in many visual interpretations. Since the facial parts separation in a human face is well-known in human anthropometry, it is possible to find the distance of a human focus from a single image frame with good accuracy provided an understanding of the camera's field-of-view. The research community in today's fast-paced technological environment wants more realistic representations, thus 3D representations of 2D images are becoming increasingly important. These methods are categorized into the following primary categories based on their applications.

1) FEATURE EXTRACTION METHODS

The expressions on people's faces reveal information about individuals. Faces identify people, and one may infer how others are feeling from their expressions. Face feature extraction can help in the improvement of face depth maps tasks. In the realm of computer vision, facial feature depth estimation and 3D reconstruction are popular topics. In computer vision-related applications such as detection and recognition, especially under shifting posture lighting, and expression, 3D information gives significant benefits in overcoming difficulties associated with 2D images (PIE) [14]. Methods have been shown in the SoA to be a potential solution to several of problems in facial depth maps [20]–[25].

2) FEATURE FUSION METHODS

Feature fusion offers a full description of image features' rich internal information, and following dimensionality reduction, compact representations of integrated features can be obtained, resulting in decreased computational complexity and better performance of facial depth maps. 3D reconstruction helps in the resolution of difficulties in 2D images as well as the improvement in performance in a variety of tasks. Several approaches have been offered in the last few years [26]–[34] for facial depth estimation tasks.

3) IMAGE PROCESSING FILTRATION METHODS

For the successful application of depth information, quality is critical. Visually undesirable rendered views are frequently produced when a depth map is distorted by large featureless artefacts. A robust depth image post-filtering technique should be considered for further 3D video transmission. Filtering of depth maps has primarily been studied from the viewpoint of increasing resolution [35]–[37]. There are a variety of post-processing techniques for restoring natural images [38]. Filtering algorithms included Gaussian smoothing and the H.264 in-loop deblocking filter [39], as well as a local polynomial approximation (LPA) [40] and bilateral filtering [41], which use edge-preserving structure information from the colour channel to refine rough depth maps [42].

TABLE 1. Properties of feature, fusion, and image processing filtration methods.

Method Category	Methods	Strategy	Category	Descriptions of the main block	Uses
Feature Extraction	[14]	Depth From Shading,	DL	Light-Field Angular Function	Depth Maps
	[20]	Defocus	DL	Adversarial Networks	Depth Maps
	[22]	Face Depth CNN	ML	Surface Normal Direction	Reconstructions
	[23]	Recovering Facial Shape	ML	Symmetric Self-Ratio Images	Reconstructions
	[25]	Shape-From-Shading	DL	Feature Extractor	Object Recognition
		From Depth Maps CNN			
Feature Fusion	[26]	Face Depth CNN	DL	Single Reference Face Shape	3D Face Reconstruction
	[27]	Face Depth CNN	DL	Multi-Level Feature Fusion	3D Reconstruction
	[28]	Autoencoder	DL	Stacked Contractive Autoencoder	Learning 3D Faces
	[29]	Single Facial Depth Map	DL	Multi-Level Feature Fusion	Refinement
	[30]	Face From Depth	DL	Feature Fusion Extractor	Driver Pose Estimation
	[31]	Face From Depth	DL	Feature Fusion Extractor	Image Super-Resolution
	[32]	Pose	DL	Multi-Level Feature Fusion	Pose Estimation
	[33]	3D Blendshape	DL	Feature Fusion Extractor	Facial Expression
	[34]	Learning Feature	ML	Multi-Level Feature Fusion	Recognition
					Aggregation
Image Processing Filtration	[36]	Learning Feature	ML	Joint Bilateral	Upsampling
	[37]	Depth	ML	Multistep Joint Bilateral	Depth Upsampling
	[38]	Pointwise Shape-Adaptive	ML	High-Quality Filtration	Denoising And Deblocking
	[39]	Pointwise Shape-Adaptive	ML	Filters	Deblocking
	[40]	Local Approximation	ML	High-Quality Filtration	Signal And Image Processing
	[41]	For Gray And Color Images	DL	Bilateral Filtering	Signal And Image Processing
	[42]	Fused Deep Representation	DL	Light Field	Face Recognition

Table 1 shows the corresponding methods categorized into feature extraction, feature fusion, and image processing filtration with their respective use cases and strategies involved.

a: FACIAL DEPTH IN 3D FACE RECOGNITION

Face recognition (FR) has been used for human identification for ages. With the advances of deep neural networks (DNNs), both face identification (one-to-many) and face verification (one-to-one) have achieved state-of-the-art results. Despite these advances, there are still a few limitations due to external conditions like viewing angles, human appearances like facial expressions, occlusions, scene lightings. To overcome these factors researchers, use other modalities like depth and surface normal. The availability of low-cost RGB-D consumer level sensors like Microsoft Kinect and Intel Real Sense which simultaneously capture depth data of the scene and the colour intensity make these multimodal data more accessible. Depth information can be very useful in FR because it helps to retrieve geometric information of the face in the form of dense 3D points. RGB-D FR can be categorized broadly into two classes – handcrafted feature-based method and deep learning-based methods. Table 2 shows the corresponding details of the listed methods for this subsection.

B. FACIAL DEPTH FROM STEREO AND MULTI-VIEW

Using two or more cameras, depth can be derived from stereo or multi-view. A process known as stereo matching is used to produce this map. The primary notion is that triangulation and stereo matching can be used to estimate depth in a variety of applications, including object grasping, collision avoidance, broadcasting, robotic navigation, and multimedia. The most frequently used methods for measuring face depth from stereo methods are designed on fitting the computed depth to a generalized 3D model [49]–[51]. For facial depth estimation, a passive stereo system for 3D human face reconstruction and recognition at a distance method is introduced [52]. Using a Kinect camera and a face detection algorithm, a method was able to reliably locate the human head and estimate head posture. To locate the detailed facial characteristics, a depth AAM algorithm is designed [53]. In a passive stereo vision system, a method for estimating facial depth is introduced. The method relies on the fast creation of facial disparity maps, which does not necessitate the use of expensive instruments or generic face models. It entails including face attributes in the disparity estimate process to improve 3D face reconstruction [54].

The primary drawbacks of these approaches are the long processing times associated with the fitting phase (due to the high computational complexity) and the need for human setup, as seen in [51]. Another drawback of these approaches is that the generated faces resemble the generic model rather than their model. It's also particularly sensitive to noise because it calculates curves using the second derivative.

C. FACIAL DEPTH FROM 2D, MONOCULAR IMAGES

The monocular depth estimation method uses only a single RGB image as input to predict the depth value of each pixel or infer depth information. The following methods use a monocular depth strategy. Monocular depth maps are simple to set up, especially when it comes to camera calibration, and only require a single image to estimate depth. It can also

TABLE 2. Properties of facial recognition depth maps methods.

Methods	Feature Type	Features extracted	Strategy	Method Category	Descriptions of the main block	Uses
[43]	Geometric	Histogram Of Oriented Gradient (HOG)	Random Decision Forest (RDF) Classifier	Feature Extraction DL	Entropy Map	Recognition
[44]	Geometric	Local Binary Patterns (LBP	Iterative Closest Point (ICP) And	Feature Extraction DL	Discriminant Color Space (DCS)	Depth Maps
[45]	Geometric	Signed Distance Function (SDF)	ICP	Feature Extraction ML	3D Face Model	Depth Maps
[46]	Statistical	Feature Space	CNN	Feature Fusion DL	Autoencoder	Depth Maps
[47]	Spatial	Feature Space	Single Facial Depth	Feature Fusion	CNN VGG	Depth Maps & Recognition
[48]	Spatial and Geometric	Feature Space	Face Recognition Accuracy	Feature Extraction	Surface Normal, Point Cloud;	Recognition & Depth Maps

give a variety of monocular visual cues, such as gradients and texture variations, colour, and defocus, that have previously been underutilized in such systems and can be used even in texture fewer areas. Table 3 shows the corresponding details of the listed methods from this section.

D. FACIAL DEPTH THROUGH DOMAIN TRANSLATION

The domain translation which is also known as image translation requires learning a parametric mapping function between two separate domains. Per-pixel classification or regression issues are frequently used to solve image-to-image translation challenges [48]–[62]. Borghi *et al.* [30], [51] suggested a method for computing the appearance of a face based on a standard CNN that combines characteristics of autoencoders and fully connected convolutional networks (FCN). Several recent studies have investigated the image-to-image translation problem by developing a mapping between two frames using conditional generative adversarial networks [52], [63]. Authors in [53] and [64], proposed an approach with the pix2pix model, which synthesizes images from semantic labelling and then reconstructs objects from edges and colourizes images. Aissaoui *et al.* [54], [65] provided a framework of linked GANs that can synthesize pairs of similar images in two separate contexts. This research also focuses on the domain translation problem to create visually attractive facial depth maps with sufficient discriminative information for face recognition.

The authors [66] present a novel framework for learning (1) RGB face parsing, (2) depth face parsing, and (3) RGB-to-depth domain translation together for facial depth maps. In [67], the authors suggest a new Deterministic Conditional GAN that is efficient for face-to-face translation from depth to RGB and is trained on labelled RGB-D face datasets. Whereas the network cannot reconstruct the exact somatic attributes of unknown focus on the individual, it can reconstruct plausible faces which is sufficient for use in various pattern recognition applications. In [68] a method proposes face from depth for head pose estimation on depth images for estimating head and shoulder pose based solely on depth images to create a complete end-to-end system. The proposed method also incorporates head detection and a localization module for facial depth estimation.

E. FACIAL DEPTH MAP DENOISING

Two forms of noise which include holes and spikes impact the depth data generated by the face reconstruction process. Pixels with unknown depth values are referred to as holes. During the disparity estimation procedure, the disparity values for these pixels are set to zero. They arise when there is an obstruction or poor light. Spikes are pixels having an incorrect depth estimation. They are mostly caused by incorrect matching and occur inhomogeneous areas where pixels have similar intensity values.

Various approaches for face depth map de-noising have been presented in the literature. These methods are divided into two categories: global and local. To eliminate spikes and fill holes, global approaches apply noise reduction filters to the hole depth image. For this, the median filter is frequently used. Authors in [69] and [70], proposed a Gaussian filter method that works to soften the data and eliminate spikes in the z-coordinate. To eliminate spikes, fill tiny gaps, and smooth the data, the authors in [71] utilized three median filters with different variances. For minor noises, these types of filters can produce optimal results. However, if the noisy region is big, these filters will not be able to remove the noise; instead, they will just modify the pixel values by their surrounding pixels.

In [49] by processing the data row by row, with the first and last non-zero pixels in each row being chosen by a sweep of the depth images. This procedure is continued until no

TABLE 3. Properties of facial depth from 2D monocular images methods.

Methods	Feature Type	Features extracted	Strategy	Method Category	Descriptions of the main block	Uses
[26]	Geometric	Single Reference Face Shape	Constrained Independent Component Analysis	Feature Extraction DL	3D Face Model	3D Face Reconstruction
[9]	Spatial & Geometric	Constrained Independent Component Analysis	The Rotation and Translation Process	Feature Extraction DL	Discriminant Color Space (DCS)	3D Face Reconstruction
[7]	Geometric	Similarity Transform & Feature Space	Deep Learning	Feature Extraction	3D Face Model	Depth Maps & 3D Face Reconstruction
[55]	Statistical	End-To-End Learning	Uses Single-View Depth and Multi-View Pose Networks	Feature Fusion	CNN Models Combined	Depth Maps
[56]	Spatial & Geometric	Canonical Correlation Analysis Surface Depth.	Surface Depth	Feature Extraction	Face Color Texture And Surface Depth	Face Depth Maps
[57]	Spatial & Geometric	Feature Points, Feature Space	Feature Points Similarity Analysis	Feature Extraction DL	Extracted To Form The 2D-3D	3D Face Reconstruction
[58]	Geometric	Recovering The Depth	Uses A Cascaded FCN And CNN Architecture	Feature Extraction	CNN Models Combined	Face Depth Estimation
[59]	Spatial & Geometric	Feature Space	Uses A Combination of Loss Function	Feature Extraction	CNN Encoder-Decoder	Face Depth Estimation

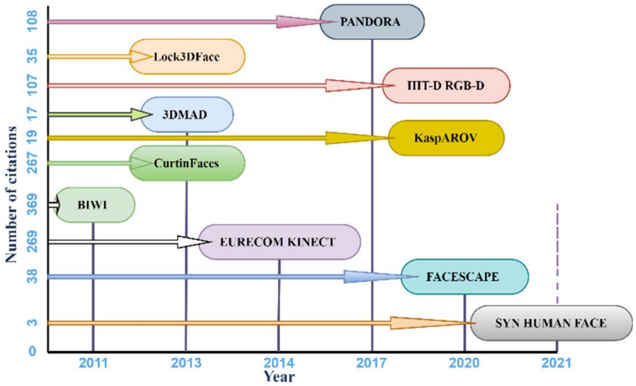


FIGURE 1. The number of facial depth datasets that are publically available each year.

more pixels are produced. The filling process usually involves utilizing an interpolation technique or a local median filter after determining the hole's boundaries. This method is more accurate than the global method since it just processes noises and leaves the non-noisy data alone. Since holes have a known value (zero or undefined), it can only handle those; spikes, on the other hand, have a random value, therefore it can't be used to eliminate them.

The authors [72] suggested an edge-guided deep neural network for the super-resolution of a single facial depth map. It is divided into two sub-networks: edge prediction and depth reconstruction. The edge prediction sub-network generates an edge guidance map that is used to guide the depth reconstruction sub-network in recovering sharp edges and fine constructions. Jovanov *et al.* [73] proposes a time-of-flight depth camera-specific wavelet-based depth video denoising approach based on multi hypothesis motion estimation for facial depth maps. In [74] authors proposed a method and system for super-solving and recovering the facial depth maps. The main idea of this approach is to use a learning-based technique to gather reliable face priors from a high-quality facial depth map to improve the depth images.

III. PUBLICLY AVAILABLE FACIAL DEPTH ESTIMATION DATASETS AND LOSS FUNCTIONS

This section provides an overview of the most commonly used facial image depth datasets, including their respective descriptions in tabular form.

There are several useful datasets available for training depth estimation methods both multi-view and monocular images of human faces. The collection's general data contains information on the number of objects, scenarios, and RGB and depth images. Among the numerous types of data contained within every dataset, the ground truth contains depth, mesh, cameras trajectories, videos, positions, point cloud, semantics label, trajectories, and dense multi-class labelling. As the field of face image depth estimation research grows in popularity, more work is being put into creating higher and additional informative depth maps datasets. Fig. 1 shows the number of new publicly available facial depth maps datasets and their corresponding number of citations becoming available each year over the period for the last ten (10) years. Table 4-6 tabulates a comparison analysis for the data existing in each dataset.

A. FACIAL AND POSE DEPTH DATASETS

The depth camera sensor should be capable of faster human-skeletal tracking in addition to being a low-cost camera sensor that outputs both RGB and depth information. This kind of tracking can provide the precise position of human body joints throughout a period, making comprehensive human behaviour investigations easier and quicker. As a consequence, there has been a lot of interest in inferring human faces from depth images and synthesizing depth and RGB images. Several new facial depths maps datasets have been generated in recent years to assist in the confirmation of humanoid facemask action analysis methods. The details of these datasets are provided in the following section.

1) BIWI

This dataset [75] comprises 15K images of 20 different subjects which included 6 female subjects and 14 male subjects (4 people were recorded twice). Moreover, this dataset provides the depth image of 640 × 480 pixels resolution, the corresponding visible image of 640 × 480 pixels size, and lastly, it also offers the annotation for every image. The depth data is captured using a Kinect v1 sensor. The dataset consist of the head poses with the range of around +−75 degrees yaw and +−60 degrees pitch. The overall dataset includes the head's 3D location and rotation as the ground truth data.

2) EURECOM KINECT FACE

This dataset provides multimodal facial data of 52 subjects among which 14 are female, and 38 are male subjects. Eurecom Kinect Face dataset [76] incorporates the depth data which is acquired from Kinect v1 sensor. This data was gathered at different times in the form of two-fold intervals with an average time gap of half month. The recorded data in two different intervals provides the facial frames of each subject in nine situations with various lighting and occlusion conditions and facial expressions which include a neutral face and smiling face.

The provided data incorporates facial data with open mouth, and different occlusions such that strong illumination, eyes occlusion by wearing sunglasses, mouth occlusion by covering it with hand, face side occlusion by placing a paper. The overall dataset provides the RGB colour images, the 3D images, and the depth map which is provided in the forms of the bitmap depth image and the text file containing the actual depth levels acquired from the Kinect sensor. The dataset also incorporates six distinct manual facial landmarks positions which comprise of right and left eye, right and left corner of the mouth, the tip of the nose, and the chin.

3) PANDORA

This dataset [30] provides a total of 250K full-resolution RGB, their corresponding depth data, and their annotations are also included in this dataset. The depth data is acquired from a Kinect v2 sensor. The Pandora dataset is frequently used for various computer vision tasks such that head poses estimation, head centre localization, and shoulder pose estimation.

4) FACESCAPE

The FaceScape dataset [78] includes large-scale 3D facial models, parametric models, and multi-view images all are recorded in high-quality. The dataset also provides the subject's age and gender, as well as the camera settings configuration. The dataset is made publicly available for non-commercial research purposes. This dataset is consisting of 3D faces acquired from 938 subjects. The overall data comprises 18,760 textured 3D faces, with 20 distinct facial expressions. The dataset provides topological information in all the 3D models by processing pore-level facial geometry. For rough shapes and intricate geometry, fine 3D facial models can be expressed as a 3D morphable model, it is represented as displacement maps. A unique methodology is proposed that takes advantage of the large-scale and high-accuracy dataset by utilizing a deep neural network to extract expression-specific dynamic characteristics.

5) 3DMAD

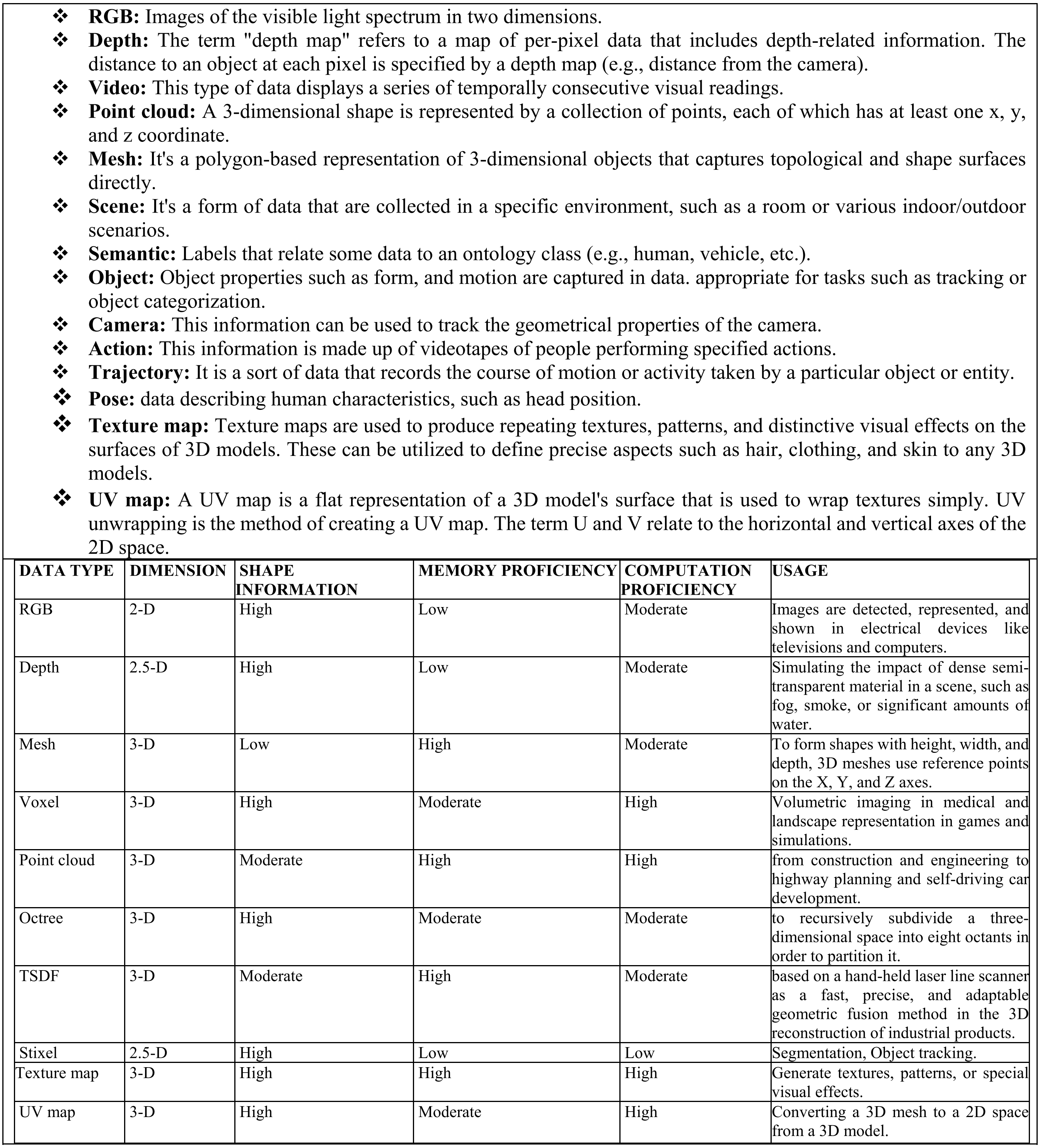
The 3D Mask Attack Database [77] (3DMAD) contains 76500 frames of 17 different subjects captured using the Kinect v1 depth sensor. Each frame is made up of a depth image with an image dimension of 640 × 480 pixels – 1 × 11 bits, a matching RGB image with an image dimension of 640 × 480 pixels – 3 × 8 bits, and precisely labelled eye locations (concerning the RGB image). Data is gathered in three distinct sessions for each subject, with each session consisting of five recordings with each recording including 300 frames. The overall data is recorded from the frontal view with neutral expression in controlled environmental conditions. The complete data is gathered in three different sessions. The first two events are for real-world samples, wherein people are recorded for two weeks. A single operator collects 3D mask attacks in the third session (attacker).

6) SYN HUMAN FACE

The SYN Human FACE [59] includes extensive high-quality 3D face models and their corresponding 2D RGB, pixel-accurate ground truth depth images. The suggested framework works as follows: In Character Creator, a collection of virtual human models is built using the real 100 head models. To generate additional data variations, the texture and morphology of the models are modified. These models are then imported to iClone for incorporating the data with five different facial expressions. The mesh, textures, and animation keyframes for the completed iClone models with individual face emotions are then exported in FBX format.

In the next phase head movement (yaw, roll, and pitch) was applied on all the models in Blender to acquire the head pose. The FBX files are then imported and scaled in the Blender world coordinate system. To replicate the real work environment, lights and cameras are included in the scene, whose properties are then adjusted accordingly. The camera sensor near and far clips have been set at 0.01 meters

TABLE 4. Comparison between data representations.



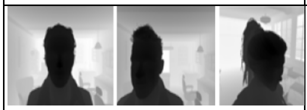
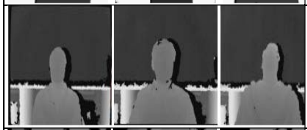
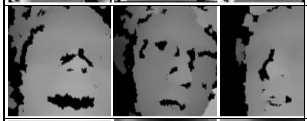
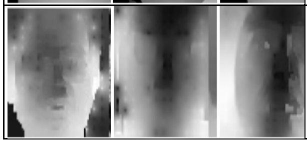
- **RGB:** Images of the visible light spectrum in two dimensions.
- **Depth:** The term "depth map" refers to a map of per-pixel data that includes depth-related information. The distance to an object at each pixel is specified by a depth map (e.g., distance from the camera).
- **Video:** This type of data displays a series of temporally consecutive visual readings.
- **Point cloud:** A 3-dimensional shape is represented by a collection of points, each of which has at least one x, y, and z coordinate.
- **Mesh:** It's a polygon-based representation of 3-dimensional objects that captures topological and shape surfaces directly.
- **Scene:** It's a form of data that are collected in a specific environment, such as a room or various indoor/outdoor scenarios.
- **Semantic:** Labels that relate some data to an ontology class (e.g., human, vehicle, etc.).
- **Object:** Object properties such as form, and motion are captured in data. appropriate for tasks such as tracking or object categorization.
- **Camera:** This information can be used to track the geometrical properties of the camera.
- **Action:** This information is made up of videotapes of people performing specified actions.
- **Trajectory:** It is a sort of data that records the course of motion or activity taken by a particular object or entity.
- **Pose:** data describing human characteristics, such as head position.
- **Texture map:** Texture maps are used to produce repeating textures, patterns, and distinctive visual effects on the surfaces of 3D models. These can be utilized to define precise aspects such as hair, clothing, and skin to any 3D models.
- **UV map:** A UV map is a flat representation of a 3D model's surface that is used to wrap textures simply. UV unwrapping is the method of creating a UV map. The term U and V relate to the horizontal and vertical axes of the 2D space.

DATA TYPE	DIMENSION	SHAPE INFORMATION	MEMORY PROFICIENCY	COMPUTATION PROFICIENCY	USAGE
RGB	2-D	High	Low	Moderate	Images are detected, represented, and shown in electrical devices like televisions and computers.
Depth	2.5-D	High	Low	Moderate	Simulating the impact of dense semi-transparent material in a scene, such as fog, smoke, or significant amounts of water.
Mesh	3-D	Low	High	Moderate	To form shapes with height, width, and depth, 3D meshes use reference points on the X, Y, and Z axes.
Voxel	3-D	High	Moderate	High	Volumetric imaging in medical and landscape representation in games and simulations.
Point cloud	3-D	Moderate	High	High	from construction and engineering to highway planning and self-driving car development.
Octree	3-D	High	Moderate	Moderate	to recursively subdivide a three-dimensional space into eight octants in order to partition it.
TSDF	3-D	Moderate	High	Moderate	based on a hand-held laser line scanner as a fast, precise, and adaptable geometric fusion method in the 3D reconstruction of industrial products.
Stixel	2.5-D	High	Low	Low	Segmentation, Object tracking.
Texture map	3-D	High	High	High	Generate textures, patterns, or special visual effects.
UV map	3-D	High	Moderate	High	Converting a 3D mesh to a 2D space from a 3D model.

and 5 meters, correspondingly. The sensor size and field of view (FOV) is set to 60 degrees and 36 mm, accordingly. The render layer's RGB and Z-pass outputs are then set up in the compositor to produce the final result. In posture mode, the head and shoulder joints are recognized, the head mesh has pivoted those bones, and the keyframes are stored to apply the rotation.

Finally, the RGB and depth images are created by rendering all of the keyframes. The matching head position (yaw, pitch, and roll) is produced using the Blender soft-

TABLE 5. Datasets of facial depth, pose, and recognition.

Examples of face images	Dataset	Labelling	Description	camera parameters	APPLICATIONS
	Biwi [75]	3d Position Of The Head And Its Rotation	People Moving Their Heads In Different Directions	Intrinsic + Extrinsic	Automatic Head Pose, Depth, Estimation, Gaze Estimation
	Eure Com Kinectv [76]	Facial Variations, Expressions, Marker Point Positions, Illumination, Occlusion	Performing Various Expressions, Poses	Intrinsic + Extrinsic, Focal Length	Face Recognition, Pose Estimation, Depth Facial Landmark Detection
	3dmad [77]	Spoofing Is Occurring, Eye Positions	3 Different Sessions For All Subjects And Each Session 5 Videos Of 300 Frames Are Captured, Neutral Expression	Intrinsic + Extrinsic	Biometric (Face) Spoofing, Facial Depth Estimation
	Pandora [30]	Head Position And Its Rotation, Features For The Face Verification	People Doing Different Poses In Front Of A Camera Poses	Intrinsic + Extrinsic	Pose, Facial Depth Estimation
	Facescape [78]	Textured 3d Face Models With Pore-Level Geometry, Expressions, Mash, Motion Map, Disparity Map, Texture	Textured 3d Faces, Captured From 938 Subjects And Each With 20 Specific Expressions	Intrinsic + Extrinsic, Focal Length	Predict Elaborate Rig Gable 3d Face Models, Facial Depth Estimation
	Syn Human Face [59]	Expression And Pose, Expressions, Meshes, 3d Position Of The Head And Its Rotation, Lighting	5 Expressions Performed By One Face, Poses, Lighting, Head And Camera Rotation, Translation	Camera Matrix Intrinsic + Extrinsic, Focal Length	Facial Depth Estimation, Pose Estimation
	Baracca Dataset [79]	Measures Of Distance, Age, Weight, Variations, Expressions	In-Car And Outside Views, Human Body Measurements	Intrinsic + Extrinsic	Thermal, Facial Depth Estimation
	Lock3DFace [80]	Changes In Facial Expression, Pose, Occlusion, And Time-Lapse	People Moving Their Heads In Different Directions	Intrinsic + Extrinsic	Pose, Facial Depth Estimation, 3D Face Analysis
	Curtinfaces [81]	Facial Variations, Expressions	Performing Various Expressions, Poses,	Camera Matrix Intrinsic + Extrinsic	Pose, Facial Depth Estimation, Face Recognition
	Iiit-D Rgb-D [82]	Head Position And Its Rotation	Performing Various Expressions Poses,	Camera Matrix Intrinsic + Extrinsic	Face Recognition, Facial Depth Estimation
	Kasparov [46]	Variations, Expressions	Poses, Lighting, Head And Camera Rotation	Intrinsic + Extrinsic	Pose, Facial Depth Estimation

ware's python module. For each frame, the RGB images are rendered with a resolution size of 640 × 480 pixels which are then stored in jpg format. Whereas the corresponding depth data is saved in a raw file (.exr format). Moreover, the head poses information for each frame is documented and stored in a text (.txt) file. The rendering process for each 2D frame nearly takes an average time duration of 26.3 seconds which is done using the Cycle Rendering Engine, provided in Blender software which is a type of physically-based path tracer for production rendering. The overall dataset consists of around 3,500k frames, with around 3.5k 2D frames per person.

The data is stored in a separate folder where each folder contains the data of 100 face models. Each face model's produced RGB images, as well as the resulting depth and head posture, are saved in three separate routes for three different backgrounds: plain, textured, and sophisticated. The synthetic dataset was used to create the sample images, which included ground truth depth images and various backdrops (basic, textured, and sophisticated).

7) BARACCA DATASET

The recent interest and growth in depth sensors have supported different methods to instinctively assess the anthropometric measurements, rather than utilising manual procedures and expensive 3D scanners. Normally, the application of depth data is limited due to the lack of depth-based public datasets including accurate anthropometric annotations. As a result, the authors [79] introduced a better dataset, Baracca, that was constructed specifically for the anthropometric measurements and vehicle perspective, including both in-cabin and outside views. This is a type of multimodal dataset that was created with synchronized depth, infrared, thermal, and RGB cameras to meet the needs of the automobile industry. The depth data is recorded using the Pico Zense DCAM710 depth sensor. The spatial resolution of the RGB sensor is 1920 × 1080 pixels, whereas the infrared/depth sensor has a resolution of 640×480 pixels. A total of 30 subjects (26 male, and 4 female) took part in the data acquisition process.

8) LOCK3DFACE

The Lock3DFace dataset [80] contains 5671 RGBD facial videos from 509 people, each with a unique facial expression, position, occluded, and moments. The database was collected throughout two periods. The very first event's neutral images are used as training examples, while the final three variations are used to create the 3 test procedures for position, occluded, and expressions. All the images from the second run, in all variants, make up a fourth validation set.

9) CURTINFACES

CurtinFaces [81] is a well-know RGBD face database that includes over 5000 co-registered RGBD images of 52 participants taken using a Microsoft Kinect. The front left, and right postures are the initial three images for each person. The remaining 49 images include 35 images with 5 different illumination variations and 7 different emotions, as well as 7 distinct positions captured with 7 facial variations. Images with sunglasses and arm occluded are also included in this collection.

10) IIIT-D RGB-D

The IIIT-D RGB-D dataset [82] includes 4605 RGBD images from 106 people collected for two periods using a Microsoft Kinect. Each participant was captured with modifications in attitude, emotion, and glasses under typical illumination conditions. The datasets which were before the procedure, which included a 5 cross-validation approach, in the tests set. The head is cropped for each image in the data.

11) KASPAROV

The KaspAROV dataset [46], which comprises automatic facial videos from 108 participants is captured by Microsoft Kinect v1 and v2 cameras. Every subject is shown in videos, each shot at a separate time. A total of 432 videos with 117,831 images are included in the dataset. Because the Kinect v2 sensor data had higher Rgbd image registration than the Kinect v1 sensor information.

B. FACIAL DEPTH ESTIMATION LOSS FUNCTIONS

On the reference depth map, deep learning-based algorithms commonly improve a regression model. The key problem for the SoA approaches in deep regression problems is determining a suitable loss function. Neural networks make use of optimization algorithms.

This error is calculated using the loss function that evaluates how well or badly the model behaves. Neural depth models have been used to estimate depth from one or many 2-D images using a variety of interesting loss functions for depth estimation challenges. This section lists the common loss functions that are used to estimate facial depth maps from one or multi 2D frame images.

1) ADVERSARIAL LOSS FUNCTION

The binary categorical cross-entropy loss function, which is used for face depth estimation in adversarial training models [20], [21], is defined as follows:

$$L_{bcc}(\mathbf{y}, r) = -\frac{1}{N}\sum\nolimits_{i=1}^{N}\left[r_i log\, y_i + (1-r_i)\, log\ (1-y_i)\right] \tag{1}$$

The discriminator output is subjected to $\mathrm{y_i = D(I_i)}$, where $\mathrm{y_i}$ is the prediction discriminator for the i-th input depth map and $\mathrm{r_i}$ is the corresponding ground truth. The goal of the generator model is to create images similar to the GT depth and the discriminator model. The mean squared error (MSE) loss function is used to achieve the first goal.

$$L_{MSE}\left(y^g, y^d\right) = \frac{1}{N}\sum\nolimits_{i=1}^{N}\|G(y_i^g) - y_i^d\|_2^2 \tag{2}$$

where $\mathrm{y^g}$ and $\mathrm{y^d}$ are the input images and the output depth map. In the second stage of the network, feed created depth images into the discriminator and use the adversarial loss on the discriminator predictions to see if the generated images can trick the discriminator model. Next, while maintaining the discriminator weights constant, back-propagate the gradients up to the generator model input and modify the generator parameters. As a result, the goal of solving the back-propagation problem is to minimize:

$$\hat{\theta}_g = arg\ min_{\theta_g}\, L_G\left(y^g, y^d\right) \tag{3}$$

where LG is a balanced sum of two components and can be defined as:

$$L_G\left(y^g, y^d\right) = \lambda \cdot L_{MSE}\left(y^g, y^d\right) + L_{bcce}\left(G\left(y^g\right), 1\right) \tag{4}$$

in which λ is a weighting parameter that controls the influence.

2) GAN LOSS FUNCTION

The loss function [20], [21] in the GAN-based facial depth model is divided into two parts: 1) Generator Loss: The generator loss is the sigmoid cross-entropy loss of the generated

TABLE 6. Publicly available depth datasets and properties for faces and poses.

DATASET	RGB	DEPTH	VIDEO	POINT-CLOUD	MESH	SCENE	SEMANTIC	OBJECT	CAMERA	ACTION	TRAJECTORY	POSE
BIWI [75]	√	√	×	×	×	√	×	×	√	×	×	√
EURECOMKINECT [76]	√	√	×	×	×	√	×	×	√	×	×	√
3DMAD [77]	√	√	×	×	×	√	×	×	√	×	×	×
PANDORA [30]	√	√	×	×	×	√	×	×	√	×	×	×
FACESCAPE [78]	√	√	×	√	√	×	×	×	×	√	√	√
SYN HUMAN FACE [59]	√	√	×	×	√	×	×	√	√	√	×	√
BARACCA DATASET [79]	√	√	×	√	×	×	×	×	√	√	×	√
LOCK3DFACE [80]	√	√	√	×	×	×	×	×	√	√	×	√
CURTINFACES [81]	√	√	×	√	×	×	×	×	√	√	×	√
IIIT-D RGB-D [82]	√	√	×	×	×	×	×	×	√	√	×	√
KASPAROV [46]	√	√	√	×	×	×	×	×	√	√	×	√

No	Dataset Name	Year	Gt	Labeling	Dimension	Objects	Subject/Type	No Images	Diversity	Annotation
1.	BIWI [75]	2011	Depth	Expression, Pose, 2D Skeleton Positions	640 × 480	Multiple	20/realistic	15K	Medium	Real RGB-D
2.	3DMAD [77]	2013	Depth	Expression, Pose, 3D Positions of The Head and its Rotation	640 × 480	Multiple	realistic	76K	Medium	Real RGB-D
3.	CURTINFACES [81]	2013	Depth, Pose	Expression, Pose, 2D Skeleton Positions	640 × 480	Multiple	52/realistic	>5K	High	Real RGB-D
4.	IIIT-D RGB-D [82]	2013	Depth, Pose	Expression, Pose	640 × 480	Multiple	106/realistic	46K	High	Real RGB-D
5.	EURECOM KINECT [76]	2014	Depth	Expression type, Pose, 2D Rotation	256 × 256	Multiple	realistic	20K	Medium	Real RGB-D
6.	LOCK3DFACE [80]	2016	Depth	Expression type, Pose, 3D Position of The Head and Its Rotation	512 × 424	Multiple	509/realistic	>6K	High	Real RGB-D
7.	KASPAROV [46]	2016	Depth	Expression type, Pose, 2D Rotation	64 × 64	Multiple	108/realistic	101K	Medium	Real RGB-D

TABLE 6. ***(Continued.)*** **Publicly available depth datasets and properties for faces and poses.**

8.	PANDORA [30]	2017	Depth	Expression, Pose, 2D Skeleton Positions	256 × 256	Multiple	20/ realistic	11K	High	Real RGB-D
9.	FACESCAPE [78]	2020	2D, 3D Landmarks, Depth	3D Position of The Head and Its Rotation	4096 × 4096	Multiple	938/Extracted	8K	High	Synthetic, 3D, RGB-B
10.	BARACCA DATASET [79]	2020	Depth	Expression, Pose	640×480	Multiple	30/ realistic	>10k	Medium	Real RGB-D
11.	SYN HUMAN FACE [59]	2021	2D, 3D Landmarks, Depth	3D Position of The Head and Its Rotation	640 × 480	Multiple	100/ Extracted	350K	High	Synthetic, 3D, RGB-B

images and an array of ones. The L1 loss function (MAE) is utilized to calculate the absolute difference between the target and generated images. This determines how similar the anticipated image is to the actual image. The following formula can be used to compute the total generator loss:

$$L_{Gen\_loss} = Gan\_loss + \lambda * L1\_loss \tag{5}$$

Here λ is set as 100.

$$MAE = \left(\frac{1}{n}\right)\sum_{i=1}^{n} |r_i - t_i| \tag{6}$$

where r_i is the prediction and t_i are the true value. 2) Discriminator Loss: The discriminator takes real images and generated images as its input. The sigmoid cross-entropy loss of the real images and an array of ones is called real loss. Then the total loss can be calculated by the summation of real loss and the generated loss:

$$T\_loss = Real\_loss + Generated\_loss \tag{7}$$

3) STRUCTURAL SIMILARITY (SSIM) LOSS

SSIM [81] is used to determine the perceived differences between the two similar images. (L_SSIM) represents the loss function for the structural similarity index measure (SSIM) and can be defined as:

$$LSSIM(r,t) = \left(\frac{1 - L_{SSIM}(r,t)}{MaxDepth}\right) \tag{8}$$

4) SCALE SHIFT-INVARIANT LOSS

For a single ag image, the scale-shift-invariant loss [81] is defined as

$$L_{SSI}(r,t) = \frac{1}{2N}\sum_{i}^{N} \rho(r,t) \tag{9}$$

where (ρ is the scale-invariant loss).

5) PRE-PIXEL SMOOTHNESS LOSS

Because image gradients commonly have depth inconsistencies, a per-pixel smoothness loss [83] is used in conjunction with the L_SL reprojection loss to make the inverse depth prediction better. The following formula is used to determine the (L_SL) loss:

$$L_{SL}(r,t) = \sum_{i}^{N} \partial_x dte^{-\partial_x(r,t)} + \partial_y dte^{-\partial_y(r,t)} \tag{10}$$

where N denotes the number of valid pixels, ∂d denotes the disparity gradient, and $\boldsymbol{e^{-\partial x,y(r,t)}}$ denotes the edges.

6) RECONSTRUCTION LOSS

When training, the network estimates disparity, and the input image is generated using the bilinear samples, utilized to recreate the image. At the local level, the bilinear sampler is completely differentiable and easily integrated into a network. A $\boldsymbol{L_{Huber}}$ and SSIM is represented as follows: which computes the inconsistencies between both the input image and the regenerated image when coupled as a photometric image reconstruction loss [19].

$$L_R(r,t) = \frac{1}{N}\sum_{i}^{N} \frac{1 - L_{SSIM}(r,t)}{2} + (1-\alpha)L_{Huber}((r,t)) \tag{11}$$

7) SCALE-INVARIANT LOSS

When training the model, depth estimation methods use the GT depth y and the predicted log depth maps. Scale-invariant loss function [81] (L_{SI}) can be represented by (L_{SI}) for the depth values and is defined as:

$$L_{SI}(r,t) = \frac{1}{N}\sum_{i}^{N} (log\,(r_i) - log\,(t_i))^2 - \frac{\lambda}{N}\left(\sum_{i}^{N} log\,(r_i) - log(t_i)\right)^2 \tag{12}$$

where λ refers to the balance factor.

8) BERHU LOSS

The OLS estimator is effective in the circumstance of checking for data with outliers or massive errors. Berhu loss, on the other hand, is designed to preserve good attributes in the face of Gaussian noise. Berhu loss function [81] (L_{Berhu}) is defined

TABLE 7. Loss functions categorized in terms of the use case applications.

Loss Function	Purpose Of Usage in Terms of Depth Estimation	Other Use Cases
Adversarial Loss Function [20], [21]	The matching feature vectors of distinct identities are linked together to expand the discriminative characteristics between them. The goal is to change the distance between two facial depth image feature vectors and predict the final depth maps.	Segmentation, 3D reconstruction, Synthetic Data generation
Gan Loss Function [22], [23]	This loss function can be used to penalize inter-subject similarities to force the estimated depth image to preserve as much subject discriminative information as feasible.	Segmentation, 3D reconstruction, Synthetic Data generation
Structural Similarity (SSIM) Loss [81]	✓ The (Structural Similarity Index) loss function is used with the BerHu loss function to use the input image structure and associated features. ✓ The perceptual difference between two similar images is measured by the SSIM loss. Details about structural loss come from relatively adjacent pixels with a deeper connection. ✓ These pixels contain vital information about the structure of the visual scene's objects.	Classification, Regression, Segmentation
Scale Shift-Invariant Loss [39]	✓ The loss function with the extra term would create a considerably smaller error because the major issue is to preserve relative depth relationships between pixels. ✓ It can also help in a diverse scene such as unknown and inconsistent scales and baselines dataset compatibility. This will allow for data to be trained on from a variety of sensing modalities, including stereo cameras (with potentially unknown calibration), laser scanners, and structured light sensors.	Regression, Segmentation, Stereo Depth Maps
Pre-Pixel Smoothness Loss	✓ This loss function estimates the similarity between the actual and predicted depth map. ✓ It also benefits the estimated depth-perceptual map's quality.	Regression, Segmentation, Stereo Depth Maps
Reconstruction Loss [19]	This loss function can be used to make the projected left-view disparity map equal to the projected right-view disparity map, resulting in more realistic disparity maps.	Segmentation, 3D reconstruction, Synthetic Data generation
Scale-Invariant Loss [39]	✓ Regardless of the absolute global size, scale-invariant loss helps in the measurement of relationships between points in the scene. ✓ The average deviation between each pixel depth prediction and the ground truth depth is all that is measured.	Regression, Segmentation, Stereo Depth Maps
Berhu Loss [40]	✓ BerHu Loss has an advantage since it uses MSE (or L2) loss to give pixels with greater residuals more weight. At the same time, it allows smaller residuals to have a larger effect on gradients than MAE loss. ✓ BerHu's loss function simply combines MAE and MSE, enhancing the whole training process and resulting in more smooth and accurate depth predictions.	Regression, Segmentation, Stereo Depth Maps
Huber Loss [40]	✓ By balancing the MSE and MAE together, the Huber Loss provides the best of both worlds. ✓ It is less sensitive to outliers in data and can predict more accurate depth maps.	Regression, Segmentation, Stereo Depth Maps

as:

$$L_{Berhu}(r,t) = \begin{cases} (r_i - t_i) & \textit{if } (r_i - t_i) \leq c, \\ \dfrac{(r_i - t_i)^2 + c^2}{2c} & \textit{if } (r_i - t_i) > c, \end{cases} \tag{13}$$

where r_i, t_i are ground truth and predicted depth maps.

9) HUBER LOSS

MSE is thought to be better at detecting outliers in a dataset, but MAE is expected to be better at preventing them. Data that appear to be outliers, on the other hand, should not be studied, and those points must not be assigned much weight. As a result, the Huber loss function [81] (L_Huber) is defined as:

$$L_{Huber}(r,t) = \begin{cases} (r_i - t_i) & \textit{if } (r_i - t_i) \geq c, \\ \dfrac{(r_i - t_i)^2 + c^2}{2c} & \textit{if } (r_i - t_i) < c, \end{cases} \tag{14}$$

where $\mathrm{r_i}$, $\mathrm{t_i}$ are ground truth and predicted depth maps.

Table 7 shows the loss function categorized according to their use in depth estimation and their respective use case applications.

IV. IMPLEMENTATION DETAILS OF NEURAL DEPTH ESTIMATION NETWORKS

Convolutional neural networks (CNN) are the form of a learning algorithm for data processing with a uniform grid, such as images, that is intended to acquire provides scalable features from low- to high-level structures efficiently and adaptively. Convolution, pooling, and fully connected layers are the three types of layers (or building blocks) that make up CNNs. Convolution and pooling layers are the initial layers that extract features, while the third, a fully connected layer, transmits these characteristics into the final output, such as classification or multiple regression analysis. A convolution layer is an important part of CNN, which is made up of a stack of mathematical computations like convolution, which is a specific sort of linear operation. Because a feature can appear everywhere in a digital image, image pixels are saved in a two-dimensional (2D) grid, i.e., an array of numbers and a small grid of parameters called the kernel, and an optimizable feature extractor, is implemented at every image position, CNNs are extremely efficient for image analysis. Features extracted can evolve hierarchical structures and progressively

TABLE 8. Performance evaluation of monocular depth estimation based deep learning models on IIIT-D RGB-D [82], KASPAROV [46], CURTIN FACES [81], and LOCK3DFACE [80].

REFERENCE	YEAR	NETWORK	DATASETS	PARAMETERS	LAYERS	INPUT/OUTPUT	ACCURACY %
[46]	2016	AUTOENCODER	IIIT-D RGB-D [82]	47M	CNN, FC, SOFTMAX	RGB/DEPTH	98.7
[84]	2014	VGG-16	KASPAROV [46]	32M	CNN, FC, SOFTMAX	RGB/DEPTH	94.4
[85]	2016	RESNET-50	IIIT-D RGB-D [82]	68M	CNN, FC, SOFTMAX	RGB/DEPTH	95.8
[86]	2017	SE-RESNET-50	CURTIN FACES [81]	86M	CNN, FC, SOFTMAX	RGB/DEPTH	97.8
[58]	2018	INCEPTION-V2	LOCK3DFACE [80]	73M	CNN, FC, SOFTMAX	RGB/DEPTH	71.7
[47]	2020	VGG + DEPTH	IIIT-D RGB-D [82]	84M	CNN, FC, SOFTMAX	RGB/DEPTH	99.6

more complicated as one layer passes its results into the next layer. Training is the process of adjusting parameters such as kernels to reduce the disparity between outputs and ground truth labels using optimization algorithms like backpropagation and gradient descent. Fig. 2 illustrates the comprehensive implementation details.

The performance of 2D facial depth estimation has been greatly enhanced because of the use of Deep Learning CNNs. Facial depth maps are learned directly from 2D RGB-D facial images by training deep neural networks on large datasets. Different deep learning models (i.e; VGG, Autoencoder, ResNet, encoder-decoder, inception, DenseNet) are used for facial depth maps which are trained on 2D face depth images. These models typically consist of CNN, FC, SoftMax layers followed by an appropriate loss function that can minimize the errors of the training networks. Weights of the networks are mostly randomly initialized. The datasets can be augmented in several ways (pose augmentation, resolution, transformation, rotation, cropping, and flipping) using a range of images to enlarge training datasets and can achieve better accuracy. Table 8, shows some comparison analysis of the deep learning-based models for facial depth estimation on iiit-d rgb-d [82], kasparov [46], curtin faces [81] and lock3dface [80] datasets. Note that we were unable to compare other qualitative evaluation metrics mentioned in Table 8 due to technical difficulties with publicly available codes and a lack of instructions for these methods listed in Table 8, and the accuracy results are obtained from their related articles. A CNN-based system has three major components, a training phase, data pre-processing, and model design. To train the model, deep learning-based techniques usually require a significant number of datasets. In CNN-based facial depth maps research, a shortage of large-scale realistic face depth datasets remains an outstanding topic. Because CNN has a lower tolerance for pose changes, suitable data preparation or synthetic data can enhance accuracy before transmitting the data to the model. In addition, selecting an appropriate CNN and loss function are critical.

V. EVALUATION METRICS FOR FACIAL DEPTH ESTIMATION

The most used quantitative metrics for evaluating the performance of monocular facial depth estimation methods are provided in Table 9. These are not limited to 8 metrics, however, most of the published articles used these quantitative metrics to analyze the performance of the trained depth estimation models.

VI. FACIAL DEPTH ESTIMATION MODEL

Many consumer applications including robotics, augmented reality and advanced driving monitoring systems can benefit from facial depth estimation neural depth networks from single images. A methodology for creating depth maps from

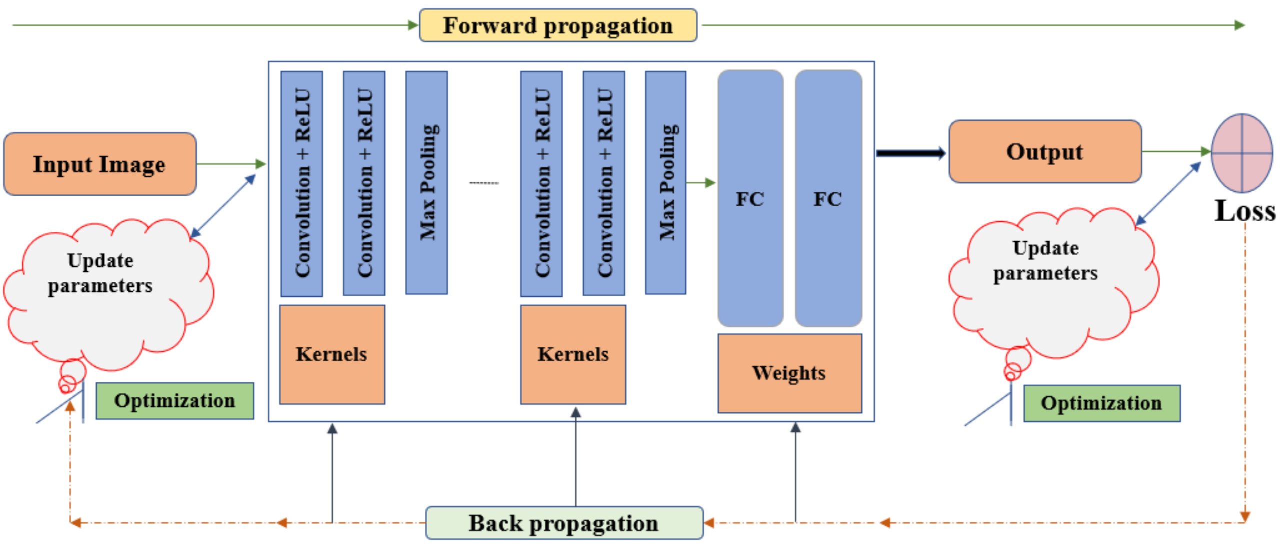


FIGURE 2. A look at the design of a CNN and how it's trained for facial depth estimation. Convolution layers, pooling layers (e.g., max-pooling), and fully connected (FC) layers are the building components that make up a CNN. The success of a model with certain kernels and weights is evaluated using a loss function and forward propagation on a training dataset, and learning parameters, such as kernels and weights, are adjusted using the gradient descent process. The term "corrected linear unit" refers to a linear unit that has been rectified.

TABLE 9. Quantitative metrics used for performance evaluation of monocular facial depth estimation.

S.No	Quantitative Metrics Name	Formula
1	AbsRel	$\frac{1}{N}\sum\frac{\|d_i-d_i^*\|}{d_i}$
2	RMSE	$\sqrt{\frac{1}{N}\sum\|d_i-d_i^*\|^2}$
3	RMSE (log)	$\sqrt{\frac{1}{N}\sum\|\log d_i-\log d_i^*\|^2}$
4	SqRel	$\frac{1}{N}\sum\frac{\|d_i-d_i^*\|^2}{d_i}$
5	Accuracies	% of $d_i\max(d_i/g_i)=\delta\ thr$
6	L1	$\sum_{i=1}^{n}\|y_{\text{true}}-y_{\text{predicted}}\|$
7	L2	$\sum_{i=1}^{n}\left(y_{\text{true}}-y_{\text{predicted}}\right)^2$
8	NRMSE	$\sqrt{\frac{1}{N}\sum\frac{\|d_i-d_i^*\|}{d_i^*}}$
where $\boldsymbol{d_i}$ and $\boldsymbol{d_i^*}$ are the ground truth and predicted depth at pixel $\boldsymbol{i}$ and $\boldsymbol{N}$ is the total number of pixels.		

single images of human faces is presented in this section, which utilizes the source face depth and corresponding ground truth depth using neural networks.

Existing facial depth map algorithms may produce depth maps with comparable accuracy, but they suffer from difficulties such as missing values and depth similarities, which result in holes in depth images. As an alternative, the model used in this study automates the collection of optimal parameters, reducing model complexity during the training process for facial depth estimation.

A recent SoA LapDepth [68] model is chosen to accomplish high-quality facial depth estimation from a single 2D frame. By applying the Laplacian pyramid-based decomposition technique to the decoding process, the suggested method intends to successfully restore local details (i.e., depth boundaries) as well as the global layout of the depth map. The depth residual including local details, which suitably describe depth attributes of different scale-spaces, is created using Laplacian residuals of the input colour image guidance encoded features. To improve the efficiency of this decoding process, the authors [87] introduce weight standardization to the pre-activation convolution block, which greatly helps in estimating depth residuals. First, describe the overall architecture of the proposed decoder for monocular facial depth estimation in this section. The entire decoding procedure will then be detailed, including the influence of weight standardization. Finally, the loss functions utilized to train the model architecture are discussed.

A. ARCHITECTURE DETAILS

The proposed neural depth network for single image facial depth maps mechanism is provided in this section, as well as the suggested loss function for improving the training process over the training data.

1) ENCODER MODEL

The proposed method's general architecture is demonstrated in Fig. 3 [87]. The suggested decoder for restoring depth residuals is connected to the pre-trained encoder in the network. ResNext10 [56] is used in the encoder phase, which has been pre-trained for image classification. The input colour image is compressed as latent information using densely layered convolution blocks on the encoder. The spatial size of such features shrinks to a fraction of the original resolution, but they compactly contain the colour-depth relationship in the embedding space, which is learned from various scene geometries. For the convolution block of the encoder, the authors utilize the Dense ASPP approach [88] with four dilation rates of 3, 6, 12, and 18 to extract more dense contextual information.

The suggested decoder is separated into many Laplacian pyramid branches. One branch, which is in charge of the Laplacian pyramid's topmost level, undertakes decoding work to restore the depth map's global layout. The depth residuals are generated by other branches using latent features led by Laplacian residuals of the input colour image at the matching scale. Using point-wise addition, this depth residual is gradually integrated with the middle depth map, which is the result of the higher level of the Laplacian pyramid. The decoding technique is based on a five-level Laplacian pyramid. All convolution layers in the decoder have a filter size of 3 × 3.

B. DECODER MODEL

The laplacian residual of the input colour image is derived in the first phase. For all scaling methods in the suggested methodology, downsampling the initial input image, upsampling, and bilinear interpolation are used. Concatenated features are input into layered convolution blocks, and the output is added pixel-by-pixel. The one-channel output, which is made up of stacked convolution blocks, has the same spatial resolution as the input colour image. It's important to note that input guides the decoding process to precisely restore local characteristics of various size areas, which aids in revealing depth boundaries without distortions. Finally, starting at the top of the Laplacian pyramid, the depth map is gradually recreated. The weight standardization in the pre-activation convolution block, which is the core module of the decoder, is made to produce the decoding process for monocular facial depth estimation more effectively. Because the depth map is reconstructed using an iterative accumulation of depth residuals, it is preferable for the projected depth residual to have a balancing of negative and positive values to estimate depth information reliably and accurately. During backpropagation, which is calculated from each layer of the laplacian pyramid, the decoder is capable of improving the flow of gradient by normalizing them. This is preferable for maintaining the colour-to-depth translation's stability based on residual information. The procedure is anticipated to be able to effectively understand the important connection between colour and depth values for facial images by combining this benefit with the Laplacian pyramid-based decomposition technique.

C. LOSS FUNCTION

The facial depth estimation task's final goal is to find a function that predicts the depth from an input image. ($\boldsymbol{L_{silog}}$) is the most common loss function that is found in the literature more helpful for depth estimation, The network's trainable parameters are tuned based on the loss function, which employs properly scaling the loss function's range can improve converging and training outputs while putting a stronger focus λ on decreasing error variance, leading in a Silog loss function [89]. ($\boldsymbol{L_{silog}}$) is defined:

$$L_{si}(d_i, d_i^*) = \frac{1}{N}\sum\nolimits_i^N (log(d_i) - log(d_i^*))^2 - \frac{\lambda}{N}\Big(\sum\nolimits_i^N log(d_i) - log(d_i^*)\Big)^2 \qquad (15)$$

where λ is the balance factor and $\boldsymbol{N}$ is the number of pixels.

By rewriting the equation. 15:

$$L_{silog}(d_i, d_i^*) = \frac{1}{N}\sum\nolimits_i^N (log(d_i) - log(d_i^*)) - \frac{1}{N}\sum\nolimits_N^i (d_i - d_i^*)^2 + (1-\lambda)\frac{1}{N}\sum\nolimits_N^i (d_i - d_i^*)^2 \qquad (16)$$

In log space, the combined Silog loss is defined as:

$$L_{silog}(d_i, d_i^*) = \alpha\sqrt{L_{silog}(d_i, d_i^*)} \qquad (17)$$

VII. EXPERIMENTAL RESULTS

The experimental results are presented in this section show how well the proposed model performs. The purpose of these experiments is to see how well synthetic facial depth data can be used to estimate facial depth estimation. A set of SoA depth estimation single image neural networks is used to analyze and compare the human facial depth estimation. Furthermore, the model is first trained on a synthetic human facial depth dataset, after which it is evaluated against four different datasets (Pandora, Eurecom Kinect Face, Biwi Kinect Head Pose, and Synthetic human face datasets) explained in section 3. After that, there is a brief comparison analysis (evaluation results of the SoA to the proposed model) is presented. The experiments show that a model trained on a large and diverse set of facial depth images, along with the appropriate training methods, produce SoA results in a variety of scenarios. The zero-shot cross-dataset transfer technique is used to demonstrate this process.

A. TRANING METHODOLOGY

The proposed approach is designed in the PyTorch tool. The suggested decoder's parameters (i.e., the network's weights) are all initialized using the approach described in [88]. The proposed decoder has group normalization in each layer, which is known to be batch size independent. The model is trained on a synthetic human facial depth dataset (described in section 3), which was divided into training and validation sets with 0.8 and 0.2 ratios for facial depth estimation. The network is trained using the Adam optimizer for 50 epochs with a batch size of 6, with power and momentum set to 0.9 and 0.999, respectively. For the encoder and decoder, the weight decaying factor is set to 0.0005 and 0. Using a polynomial

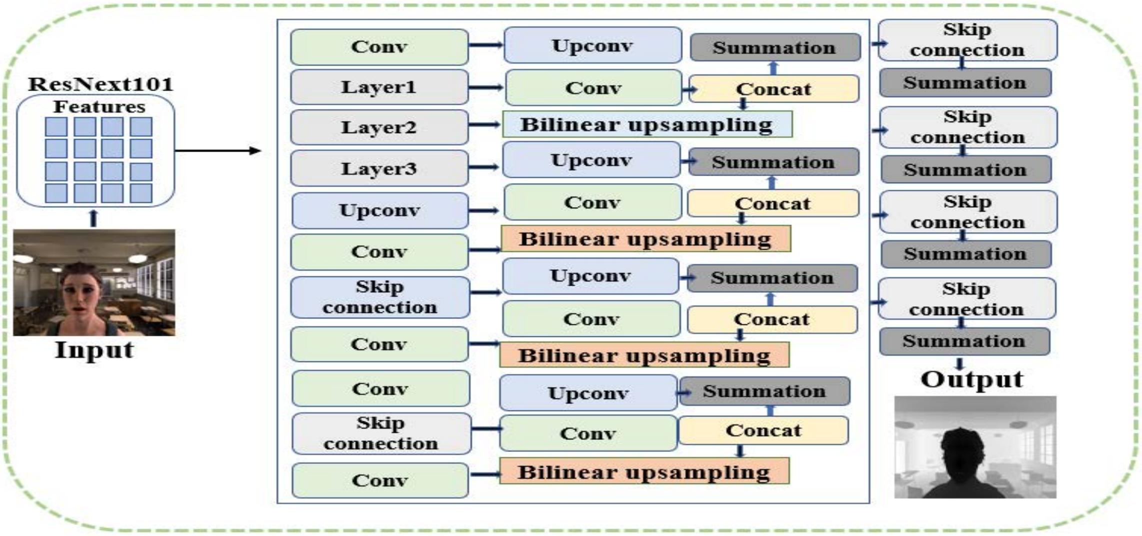


FIGURE 3. The overall architecture of the proposed method for monocular facial depth estimation.

decay with the power of 0.5, the learning rate is first set to 10^{-4} and then gradually decreased until it reaches 10^{-5}. The overall training process is conducted on a machine equipped with two TITAN 1080 GPUs, which takes a time duration of 72 hours. The model has 73M parameters and to avoid overfitting, the online data augmentation method is used in the training process. For the SYN HUMAN FACE dataset, training samples are randomly cropped to 512 × 416 pixels before being randomly rotated in the range of [3, 3] degrees. With a ratio of 0.5, input images are also horizontally flipped. Furthermore, the scale factor picked from the range of [0.9, 1.1] is used to alter the brightness, colour, and gamma values of the input colour images.

B. EXPERIMENTAL DETAILS AND RESULTS

The first phase of this subsection explains the training dataset that was used to train the neural depth model for facial depth estimation. The second part explains the testing and evaluation process used to evaluate the model's generalization performance. For evaluations, Root Mean Square Error (RMSE), log Root Mean Square Error (RMSE (log)), Absolute Relative difference (AbsRel), Square Relative error (SqRel) and Accuracies are used defined in Table 9. Four test datasets were chosen based on the diversity and accuracy of their ground truth. The model's performance is compared to existing SoA approaches in the final phase. Table 10 summarizes all of the information from this study's experiments.

TABLE 10. Information about how experiments have been conducted.

Method	LapDepth [87]
Tools/Software	PyTorch, Open3d
Training Time	72 hours
Input	512×416
Output	512×416
Type	CNN (Encoder-Decoder)
Optimizer	Adam
Learning Rate	10^{-5}
Environment	2×TITAN 1080 GPUs 2.5Ghz Python
Memory	16×2GB
Epochs	50
Parameters	73M

1) MODEL TRAINING DATASET

The synthetic human facial dataset having various variations including camera location, light position, body-pose, facial animations, scene illuminations, and pixel-accurate ground truth depth is used for training the proposed neural depth model for facial depth maps. This dataset is briefly explained in (section 3-part A subsection 6. Before conducting any experiments, the training data is processed and split into three sets: training set 80%, validation set 20%, and test set 10%, each having its ground truth depth.

2) TEST DATASETS

For comparison purposes, the zero-shot cross-dataset transfer protocol is utilized. The model was trained on a single dataset before being tested on unseen test datasets. The four datasets described in (section 3-part A) were chosen for testing and evaluation (i.e, Pandora, Eurecom Kinect Face, Biwi Kinect Head Pose, and Synthetic human face datasets).

3) MODEL PERFORMANCE EVALUATION

The performance of the facial depth estimation model LapDepth [87] is compared to the SoA models (i.e., MiDaS

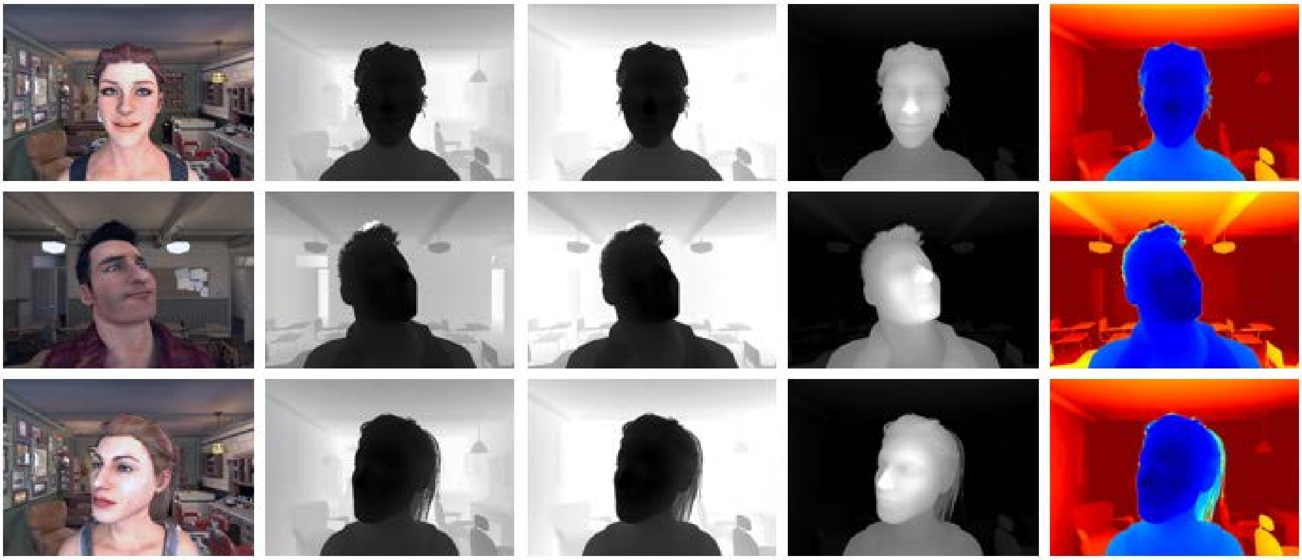


FIGURE 4. Qualitative results in a sample of the synthetic human facial test dataset that was not used for training or validation. Input RGB images, ground truth images, predicted depth images, predicted depth images (Greys), and predicted depth images are shown from left to right.

[90], DPT [91], and BTS [89]) on the synthetic human facial dataset in Fig. 4 and Table 11. All of the training and testing experiments in this work have been coded and are available on Github. The network achieves SoA results, as shown in Table 11. The proposed model qualitative results against SoA approaches are shown in Fig. 5 and Fig. 6. As shown in Fig. 5, the results demonstrated a details information and consistency, indicating that the proposed chosen approach works better at facial depth estimation. The model outperformed SoA both numerically and qualitatively in tests across a variety of real and synthetic images and set a new SoA for facial depth estimation.

In comparison to other SoA methods, the LapDepth approach performed best in terms of accuracy and depth range, according to the comparison analysis Table 11 and Fig. 6. As shown in Table 11, the network achieved 0.0281 RMSE and 0.9976 threshold accuracy on a synthetic human facial dataset (row 8). For better visualization, the results are shown in the different colour maps. Note that, predicted depth images (Greys) indicate the inverse depth map Fig 4.

As mentioned before the most commonly used quantitative metrics for evaluating the performance of trained monocular facial depth estimation methods are provided in Table 9. Based on the metrics in Table 11 i.e.; RMSE, RMSElog, SqRel, AbsRel, and accuracies one can compare and decide which method performance is better.

The model is compared with the SoA models (i.e.; MiDaS [90], DPT [91], and BTS [89]) for comparison, and the qualitative results are shown in Fig. 5. We were unable to train the techniques (i.e. MiDaS, DPT) from scratch due to unavailability of the training codes and a lack of instructions, and hence simply fine-tuned the model checkpoint for testing and validation purposes. The method BTS is initially trained on a training dataset before being put to the test on four different datasets. The suggested method has an advantage over the BTS and other SoA methods, as shown in Fig. 5. The model can recover fine details such as facial information and backgrounds since it is trained on pixel-accurate ground truth depth facial data. Pandora, Eurecom Kinect Face, and Biwi Kinect Head Pose are among the datasets that rarely capture those datils. It is difficult to learn when training neural depth networks due to a very sparse ground truth depth. It is noticed that the method LapDepth successfully preserves the facial depth information even with complicated geometries as compared to the rest of the SoA approaches. As can be seen in Fig. 6, the results show improved information and consistency, demonstrating that the works were better at depth estimation on real facial depth datasets. The network was not used for training or validation, and the method was exclusively trained on synthetic human facial depth datasets and tested on real datasets. In fig. 5, the results in the 4th column predicted depth images (Greys) indicate the inverse depth maps that is originally used by MiDaS [90]. The rest of the comparison results are respectively calculated with the same scale while predicting the depth estimation models.

VIII. DISCUSSION

The results presented in the previous section are discussed in the following section.

1. The model is trained by using only the Synthetic Human Facial Depth Dataset and evaluated against four different datasets, including the Pandora dataset, Eurecom Kinect Face dataset, Biwi Kinect Head Pose

TABLE 11. Quantitative evaluations on the SNY human face dataset [59].

No.	Methods	AbsRel	SqRel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
1.	DenseDepth-169 [92]	0.0296	0.0096	0.0373	0.0129	0.9890	0.9920	0.9981
2.	ResNet-101 [59]	0.0123	0.0210	0.0306	0.0089	0.9938	0.9965	0.9980
3.	EfficientNet-B0 [93]	0.0145	0.0280	0.0360	0.0154	0.9912	0.9934	0.9978
4.	BTS [89]	0.0165	0.0092	0.0206	0.0102	0.9830	0.9943	0.9956
5.	UNet-simple [94]	0.0103	0.0207	0.0281	0.0089	0.9960	0.9976	0.9987
6.	MiDaS [90]	0.0146	0.0204	0.03560	0.0323	0.9665	0.9902	0.9956
7.	DPT [91]	0.0156	0.0106	0.0394	0.0184	0.9567	0.9646	0.9943
8.	LapDepth [87]	0.0145	0.0041	0.0204	0.3614	0.9545	0.9857	0.99582

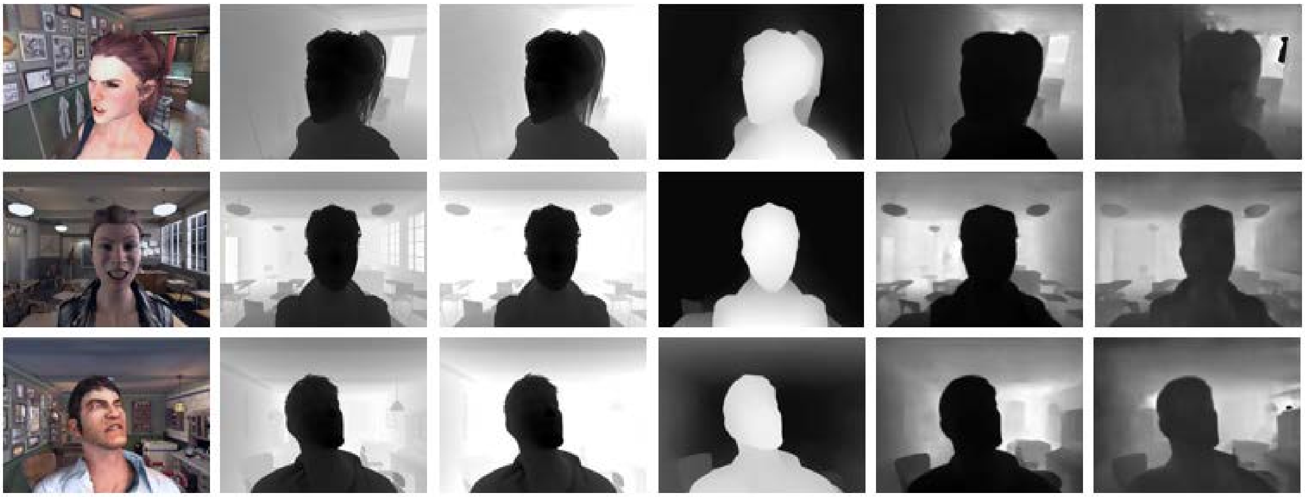


FIGURE 5. From left to right, qualitative results of facial monocular depth estimation algorithms (Input: input RGB images; GT: ground truth images; Ours: LapDepth [87], MiDaS [90], DPT [91], and BTS [89] applied to the Synthetic human facial dataset [59]).

dataset, and the test Synthetic Human Facial Depth Dataset, as well as real images, in the testing phase. The results demonstrate that the trained model outperforms the other SoA approaches MiDaS, DPT, and BTS. It is important to mention that the low size and diversity of the Pandora dataset, Eurecom Kinect Face dataset, Biwi Kinect Head Pose dataset do not perform well on the generalization performance of the studied models, as shown in Fig. 6. Furthermore, most depth GT are error-prone due to practical restrictions in data gathering. The depth GT data is particularly prone to mistakes in these datasets that make it difficult for models to learn robust facial depth information.

2. Synthetic facial data will, of course, lack the same level of detail in terms of skin features as compared to real-world image data. However, considering the numerous advantages of utilizing synthetic data to train a neural depth model, it acquires comparable accuracy to real-world data as shown in Fig. 6.
3. When the new loss function is utilized in the final set of experiments, the model outperforms SoA when the network is trained entirely on synthetic data. As a result, it is rational to assume that employing a scalable loss function and training technique helps in acquiring greater accuracy and facial depth information.
4. The model measure how effectively the created faces preserve the individual visual features of the subjects, which requires both high and low-level features to work effectively. The suggested model allows for the maximum test accuracy and outperforms the previous models that have been examined. Based on the results, the model can estimate both high-level and low-level aspects of facial depth maps, resulting in realistic and discriminative results.

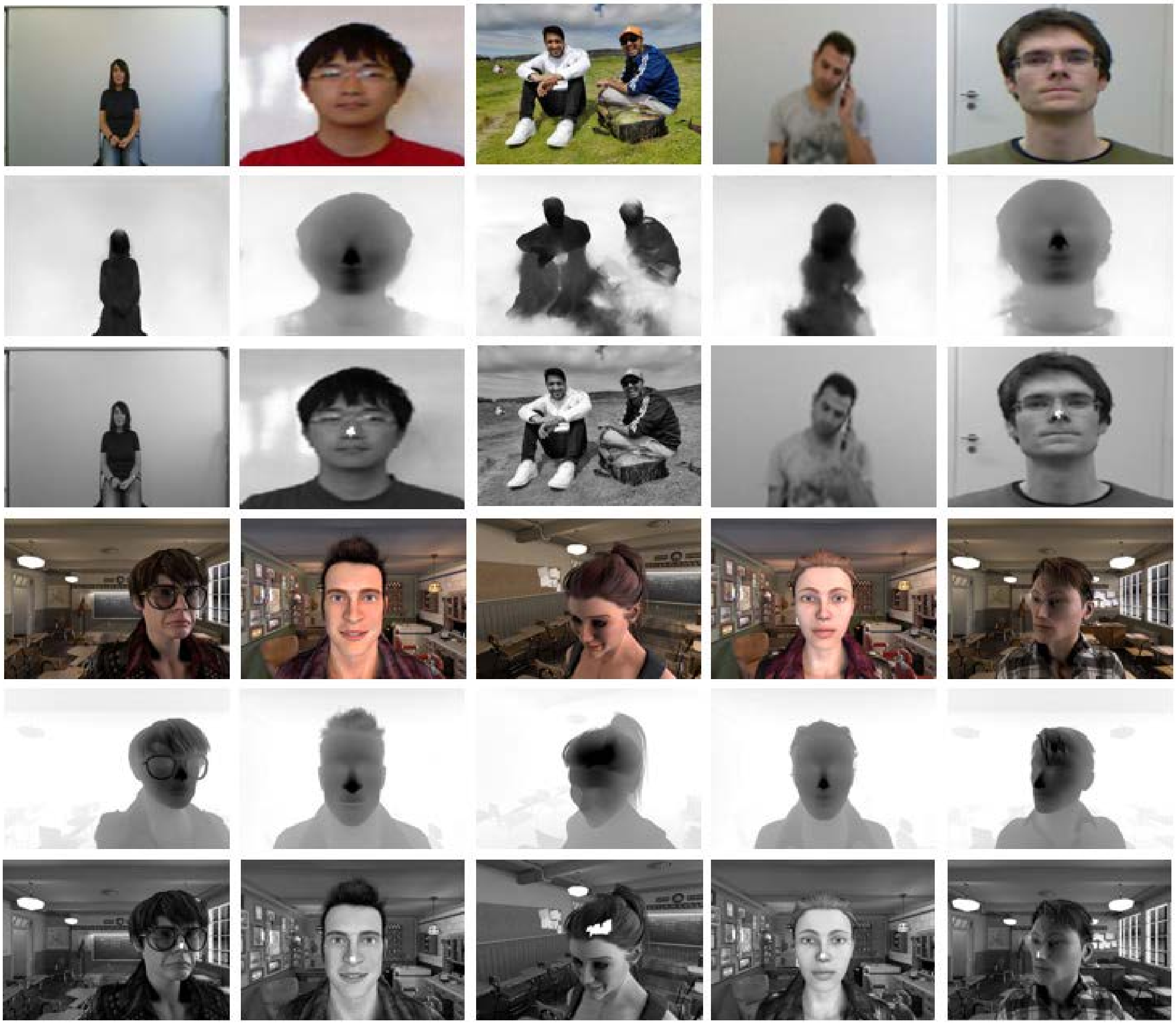


FIGURE 6. The results of a facial monocular depth estimation method's qualitative evaluation. It demonstrates how to use data from several, independent sources to estimate facial depth in a single view, despite changing and unknown depth range and scale. The method allows for broad generalization across datasets. Input images at the top. Middle: depth maps predicted by the approach provided. Bottom: corresponding point clouds as seen from a different perspective. Open3D [95] was used to render point clouds. Images from the Synthetic human facial dataset, the Pandora dataset, the Eurecom Kinect Face dataset, and the Biwi Kinect Head Pose dataset, as well as a real image of the main authors that were not seen during training.

5. Using the model predicted depth maps, as shown in Fig. 6 (row 3 and 6), the corresponding point clouds can be generated from a different perspective. Many developing visual applications require quick, direct, and exact depth information, which points clouds deliver. To localize and navigate, autonomous technologies such as robots, augmented reality devices, and self-driving cars rely on depth. In high-end smartphones, depth also enables computational photography functions like auto focus and portrait mode, which are especially useful at night when depth is difficult to obtain with traditional cameras but is readily available from a LiDAR.

IX. CONCLUSION AND FUTURE RESEARCH

This paper investigated the comprehensive details of facial depth datasets and loss functions generated in the field of computer vision for facial depth estimation problems. In various facial depth map tasks based on deep learning networks, publicly available facial depth datasets and facial depth-based loss functions have obtained robust results. The facial depth datasets are utilized in a variety of applications, including person detection and action recognition, face and pose detection, and biomedical applications. Implementation details of how neural depth networks work, as well as their associated evaluation matrices, are presented in this study. In addition to this, SoA neural architecture for facial depth estimation is proposed, along with a comparison evaluation. The proposed model outperforms current SoA techniques when tested against four different datasets. The proposed method's unique loss function helps the network in learning information aspects more robustly thus providing a detailed prediction. The training is done using synthetic human facial depth datasets, while the evaluation is done with real as well as synthetic facial images. The results prove that the proposed neural model outperforms current SoA networks, thus establishing a new benchmark for facial depth mapping and research aspects. Also, the achieved results presented in this paper can be utilized as a reference for better facial depth estimation model design and validation purposes.

Future research can be focused on developing more robust neural networks, as well as paying more attention to the newly developed facial depth datasets to obtain pixel-accurate ground truth depth maps. Because the currently available datasets have issues, particularly with realistic human faces, they can be employed in a range of real-world applications such as in-cabin driver monitoring, robotics, and 3D face reconstructions if these difficulties are addressed.

Finally, the available SoA depth estimation models can be reconsidered for the prediction of facial depth maps because they are mostly used for indoor and outdoor scene tasks and have not been extensively studied for human faces. They can also be investigated for other tasks such as single view facial recognition and surface normal prediction, 3D reconstructions, and while training on datasets both real and synthetic. The GitHub code is available online and can be found at this URL https://github.com/khan9048/LapDepth-for-Facial-depth-estimation-.

REFERENCES

[1] S.-F. Wang and S.-H. Lai, "Reconstructing 3D face model with associated expression deformation from a single face image via constructing a low-dimensional expression deformation manifold," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2115–2121, Oct. 2011.

[2] L. Spreeuwers, "Fast and accurate 3D face recognition," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 389–414, Jul. 2011.

[3] R. Lengagne, P. Fua, and O. Monga, "3D stereo reconstruction of human faces driven by differential constraints," *Image Vis. Comput.*, vol. 18, no. 4, pp. 337–343, Mar. 2000.

[4] J. Choi, G. Medioni, Y. Lin, L. Silva, O. Regina, M. Pamplona, and T. C. Faltemier, "3D face reconstruction using a single or multiple views," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3959–3962.

[5] C. K. Chow and S. Y. Yuen, "Recovering shape by shading and stereo under Lambertian shading model," *Int. J. Comput. Vis.*, vol. 85, no. 1, pp. 58–100, Oct. 2009.

[6] H.-S. Koo and K.-M. Lam, "Recovering the 3D shape and poses of face images based on the similarity transform," *Pattern Recognit. Lett.*, vol. 29, no. 6, pp. 712–723, Apr. 2008.

[7] Z. L. Sun, K. M. Lam, and Q. W. Gao, "Depth estimation of face images using the nonlinear least-squares model," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 17–30, Jan. 2013.

[8] J. Fortuna and A. M. Martinez, "Rigid structure from motion from a blind source separation perspective," *Int. J. Comput. Vis.*, vol. 88, no. 3, pp. 404–424, Jul. 2010.

[9] Z.-L. Sun and K.-M. Lam, "Depth estimation of face images based on the constrained ICA model," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 360–370, Jun. 2011.

[10] K. Konda and R. Memisevic, "Unsupervised learning of depth and motion," 2013, *arXiv:1312.3429*.

[11] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun, "Continuous Markov random fields for robust stereo estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 45–58.

[12] P. Cavestany, A. L. Rodriguez, H. Martinez-Barbera, and T. P. Breckon, "Improved 3D sparse maps for high-performance SFM with low-cost omnidirectional robots," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 4927–4931.

[13] L. Ding and G. Sharma, "Fusing structure from motion and lidar for dense accurate depth map estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1283–1287.

[14] M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, "Depth from shading, defocus, and correspondence using light-field angular coherence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1940–1948.

[15] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Opt. Eng.*, vol. 19, no. 1, 1980, Art. no. 191139.

[16] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2800–2810.

[17] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 3, Jan. 2014, pp. 2366–2374.

[18] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 740–756.

[19] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6602–6611, doi: 10.1109/CVPR.2017.699.

[20] A. T. Arslan and E. Seke, "Face depth estimation with conditional generative adversarial networks," *IEEE Access*, vol. 7, pp. 23222–23231, 2019.

[21] R. Dovgard and R. Basri, "Statistical symmetric shape from shading for 3D structure recovery of faces," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 99–113.

[22] W. A. P. Smith and E. R. Hancock, "Recovering facial shape using a statistical model of surface normal direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1914–1930, Dec. 2006.

[23] W. Y. Zhao and R. Chellappa, "Symmetric shape-from-shading using self-ratio image," *Int. J. Comput. Vis.*, vol. 45, no. 1, pp. 55–75, Oct. 2001.

[24] Q. Jin, J. Zhao, and Y. Zhang, "Facial feature extraction with a depth AAM algorithm," in *Proc. 9th Int. Conf. Fuzzy Syst. Knowl. Discovery*, May 2012, pp. 1792–1796.

[25] C. Jordan, "Feature extraction from depth maps for object recognition," Tech. Rep., 2013.

[26] I. Kemelmacher-Shlizerman and R. Basri, "3D face reconstruction from a single image using a single reference face shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 394–405, Feb. 2011.

[27] A. T. Baby, A. Andrews, A. Dinesh, A. Joseph, and V. K. Anjusree, "Face depth estimation and 3D reconstruction," in *Proc. Adv. Comput. Commun. Technol. High Perform. Appl. (ACCTHPA)*, Jul. 2020, pp. 125–132, doi: 10.1109/ACCTHPA49271.2020.9213233.

[28] J. Zhang, K. Li, Y. Liang, and N. Li, "Learning 3D faces from 2D images via stacked contractive autoencoder," *Neurocomputing*, vol. 257, pp. 67–78, Sep. 2017.

[29] F. Zhang, N. Liu, Y. Hu, and F. Duan, "MFFNet: Single facial depth map refinement using multi-level feature fusion," *Signal Process., Image Commun.*, vol. 103, Apr. 2022, Art. no. 116649.

[30] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "POSEidon: Face-from-depth for driver pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4661–4670.

[31] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[32] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[33] S. Wang, Z. Cheng, X. Deng, L. Chang, F. Duan, and K. Lu, "Leveraging 3D blendshape for facial expression recognition using CNN," *Sci. China Inf. Sci.*, vol. 63, no. 2, Feb. 2020, Art. no. 120114.

[34] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[35] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[36] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, Jul. 2007.

[37] A. K. Riemens, O. P. Gangwal, B. Barenbrug, and R.-P. Berretty, "Multi-step joint bilateral depth upsampling," *Proc. SPIE*, vol. 7257, pp. 192–203, Jan. 2009.

[38] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images," *IEEE Trans. Image Process.*, vol. 16, no. 5, pp. 1395–1411, May 2007.

[39] P. List, A. Joch, J. Lainema, G. Bjontegaard, and M. Karczewicz, "Adaptive deblocking filter," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 614–619, Jul. 2003.

[40] V. Katkovnik, K. Egiazarian, and J. Astola, *Local Approximation Techniques in Signal and Image Processing*, 2006.

[41] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vis.*, 1998, pp. 839–846.

[42] C. Angermann, M. Schwab, M. Haltmeier, C. Laubichler, and S. Jónsson, "Unsupervised single-shot depth estimation using perceptual reconstruction," 2022, *arXiv:2201.12170*.

[43] A. Sepas-Moghaddam, P. L. Correia, K. Nasrollahi, T. B. Moeslund, and F. Pereira, "Light field based face recognition via a fused deep representation," in *Proc. IEEE 28th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2018, pp. 1–6.

[44] L. Jiang, J. Zhang, and B. Deng, "Robust RGB-D face recognition using attribute-aware loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2552–2566, Oct. 2020.

[45] G. Mu, D. Huang, G. Hu, J. Sun, and Y. Wang, "Led3D: A lightweight and efficient deep approach to recognizing low-quality 3D faces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5773–5782.

[46] A. Chowdhury, S. Ghosh, R. Singh, and M. Vatsa, "RGB-D face recognition via learning-based reconstruction," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2016, pp. 1–7.

[47] H. Zhang, H. Han, J. Cui, S. Shan, and X. Chen, "RGB-D face recognition via deep complementary and common feature learning," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, May 2018, pp. 8–15.

[48] S. Pini, G. Borghi, R. Vezzani, D. Maltoni, and R. Cucchiara, "A systematic comparison of depth map representations for face recognition," *Sensors*, vol. 21, no. 3, p. 944, Jan. 2021.

[49] V. Le, H. Tang, L. Cao, and T. S. Huang, "Accurate and efficient reconstruction of 3D faces from stereo images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 4265–4268.

[50] Y. Zheng, J. Chang, Z. Zheng, and Z. Wang, "3D face reconstruction from stereo: A model based approach," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3, Sep. 2007, pp. III-65.

[51] J. R. A. Moniz, C. Beckham, S. Rajotte, S. Honari, and C. Pal, "Unsupervised depth estimation, 3D face rotation and replacement," 2018, *arXiv:1803.09202*.

[52] M. Abdelrahman, A. Ali, S. Elhabian, H. Rara, and A. A. Farag, "A passive stereo system for 3D human face reconstruction and recognition at a distance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 17–22.

[53] G. Kanojia and S. Raman, "FacialStereo: Facial depth estimation from a stereo pair," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, vol. 3, 2014, pp. 686–691.

[54] A. Aissaoui, J. Martinet, and C. Djeraba, "Rapid and accurate face depth estimation in passive stereo systems," *Multimedia Tools Appl.*, vol. 72, no. 3, pp. 2413–2438, Oct. 2014.

[55] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.

[56] M. Reiter, R. Donner, G. Langs, and H. Bischof, *Estimation of Face Depth Maps From Color Textures Using Canonical Correlation Analysis*, 2006.

[57] D. Kong, Y. Yang, Y.-X. Liu, M. Li, and H. Jia, "Effective 3D face depth estimation from a single 2D face image," in *Proc. 16th Int. Symp. Commun. Inf. Technol. (ISCIT)*, Sep. 2016, pp. 221–230.

[58] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2D face recognition via discriminative face depth estimation," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 140–147.

[59] F. Khan, S. Hussain, S. Basak, J. Lemley, and P. Corcoran, "An efficient encoder–decoder model for portrait depth estimation from single images trained on pixel-accurate synthetic data," *Neural Netw.*, vol. 142, pp. 479–491, Oct. 2021, doi: 10.1016/j.neunet.2021.07.007.

[60] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, Jul. 2016.

[61] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 577–593.

[62] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," 2016, *arXiv:1609.03126*.

[63] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[64] M. Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 469–477.

[65] D. Huang, K. Ouji, M. Ardabilian, Y. Wang, and L. Chen, "3D face recognition based on local shape patterns and sparse representation classifier," in *Proc. Int. Conf. Multimedia Modeling*, 2011, pp. 206–216.

[66] J. Lee, B. Bhattarai, and T.-K. Kim, "Face parsing from RGB and depth using cross-domain mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1501–1510.

[67] M. Fabbri, G. Borghi, F. Lanzi, R. Vezzani, S. Calderara, and R. Cucchiara, "Domain translation with conditional GANs: From depth to RGB face-to-face," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1355–1360.

[68] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara, "Face-from-depth for head pose estimation on depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 596–609, Mar. 2020.

[69] S. Berretti, A. del Bimbo, and P. Pala, "3D face recognition using isogeodesic stripes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2162–2177, Dec. 2010.

[70] Y. Wang, J. Liu, and X. Tang, "Robust 3D face recognition by local shape difference boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1858–1870, Oct. 2010.

[71] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn, "A region ensemble for 3-D face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 62–73, Mar. 2008.

[72] F. Zhang, N. Liu, L. Chang, F. Duan, and X. Deng, "Edge-guided single facial depth map super-resolution using CNN," *IET Image Process.*, vol. 14, no. 17, pp. 4708–4716, Dec. 2020.

[73] L. Jovanov, A. Pižurica, and W. Philips, "Denoising algorithm for the 3D depth map sequences based on multihypothesis motion estimation," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, pp. 1–17, Dec. 2011.

[74] S. Yang, S. Song, Q. Guo, X. Lu, and J. Liu, "Facial depth map enhancement via neighbor embedding," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1249–1254.

[75] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," in *Proc. Joint Pattern Recognit. Symp.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 6835, 2011, pp. 101–110, doi: 10.1007/978-3-642-23123-0_11.

[76] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," *IEEE Trans. Syst., Man, Cybern. A, Syst.*, vol. 44, no. 11, pp. 1534–1548, Nov. 2014.

[77] N. Erdogmus and S. Marcel, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2013, pp. 1–8.

[78] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 601–610.

[79] S. Pini, A. D'Eusanio, G. Borghi, R. Vezzani, and R. Cucchiara, "Baracca: A multimodal dataset for anthropometric measurements in automotive," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–7.

[80] J. Zhang, D. Huang, Y. Wang, and J. Sun, "Lock3DFace: A large-scale database of low-cost Kinect 3D faces," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.

[81] B. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 186–192, doi: 10.1109/WACV.2013.6475017.

[82] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, "On RGB-D face recognition using Kinect," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2013, pp. 1–6.

[83] M. Carvalho, B. L. Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "On regression losses for deep depth estimation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2915–2919.

[84] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[85] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[86] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[87] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using Laplacian pyramid-based depth residuals," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4381–4393, Nov. 2021.

[88] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.

[89] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," 2019, *arXiv:1907.10326*.

[90] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," 2019, *arXiv:1907.01341*.

[91] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.

[92] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," 2018, *arXiv:1812.11941*.

[93] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 10691–10700.

[94] F. Khan, S. Basak, and P. Corcoran, "Accurate 2D facial depth models derived from a 3D synthetic dataset," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2021, pp. 1–6.

[95] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," 2018, *arXiv:1801.09847*.

FAISAL KHAN received the bachelor's degree in mathematics from the University of Malakand, Chakdara, Pakistan, in 2015, and the master's degree in mathematics from Hazara University Mansehra, Pakistan, in 2017. He is currently pursuing the Ph.D. degree with the National University of Ireland Galway (NUIG). He also works at FotoNation/Xperi. His research focuses on deep neural networks for machine learning applications in computer vision, such as depth estimation and 3-D reconstruction.

MUHAMMAD ALI FAROOQ received the B.E. degree in electronic engineering from Iqra University, in 2012, and the M.S. degree in electrical control engineering from the National University of Sciences and Technology (NUST), in 2017. He is currently pursuing the Ph.D. degree with the National University of Ireland Galway (NUIG). His research interests include machine vision, computer vision, video analytics, and sensor fusion. He has won the prestigious H2020 European Union (EU) Scholarship and currently working with NUIG, one of the consortium partners in the Heliaus (thermal vision augmented awareness) project funded by EU.

WASEEM SHARIFF received the B.E. degree in computer science from the Nagarjuna College of Engineering and Technology (NCET), in 2019, and the M.S. degree in computer science, specializing in artificial intelligence, from the National University of Ireland Galway (NUIG), in 2020. He is currently working as a Research Assistant with the NUIG. He is associated with Heliaus (thermal vision augmented awareness) project. He is also allied with FotoNation/Xperi research team. His research interests include machine learning utilizing deep neural networks for computer vision applications, including working with visible, synthetic data, thermal data, and other bio-sensors.

SHUBHAJIT BASAK (Graduate Student Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the West Bengal University of Technology, India, in 2011, and the M.Sc. degree in computer science from the National University of Ireland Galway, Ireland, in 2018. He is currently pursuing the Ph.D. degree in computer science with the National University of Ireland Galway, Ireland. He has more than six years of experience as a software developer in the corporate world. He also works at FotoNation/Xperi. Deep learning tasks relating to computer vision are among his research interests.

PETER CORCORAN (Fellow, IEEE) is currently the Personal Chair in electronic engineering with the College of Science and Engineering, National University of Ireland Galway. He was a co-founder of many start-up firms, including FotoNation, which is now part of the Xperi Corporation's Imaging Division. He has over 600 technical publications and patents under his belt, as well as over 100 peer-reviewed journal articles, 120 international conference papers, and is a co-inventor on over 300 granted U.S. patents. For over 25 years, he has been a member of the IEEE Consumer Electronics Society. He has been named an IEEE Fellow for his contributions to digital camera technologies, particularly in-camera red-eye correction and facial recognition. He is the Founding Editor and the Editor-in-Chief of *IEEE Consumer Electronics Magazine*.

Appendix J

Report for NUIG Data Protection Office

Authors' Contribution

Contribution Criteria	*Contribution Percentage*
Research Hypothesis	PC: 80%, MAF: 20%
Experiments and Implementation	PC: 60%, MAF: 20%, WS: 20%
Background	PC: 100%
Manuscript Preparation	PC: 60%, MAF: 20%, WS: 20%

Report for NUIG Data Protection Office

Subject: Facial Recognition (FR) from Thermal Camera Data

Authors: Peter Corcoran, Muhammad Ali Farooq, Waseem Shariff

Scope of this Document

The C3 Imaging group has gathered and annotated a large dataset of thermal camera images. These are obtained from a vehicle-mounted camera around Galway city and consist of roadway scenes with pedestrians, cyclists, vehicles. The goal of this document is to explain that the thermal image data of these objects/individuals cannot be used to reconstruct a face or vehicle numberplate in sufficient detail to identify an individual or a particular vehicle.

More specifically this document addresses specific concerns regarding the potential to reconstruct facial detail with sufficient resolution to implement a useful facial recognition (FR) and thus to identify individuals within thermal image data.

Technology Background - Thermal Vs Conventional Digital Cameras

There are a number of significant differences between thermal imaging, in particular the type of thermal sensor used in the camera we have employed to gather data and conventional CMOS sensors in digital cameras. These are discussed in this section, in particular with a focus on the suitability and performance of thermal imaging when applied for facial recognition.

Thermal Imaging Technology used in this Study

The thermal image capture is performed using uncooled microbolometer sensing elements. Each 'pixel' of the image sensor is effectively an independent micro-bolometer. The current state-of-art for this sensor technology is described in this document:

- Yu L, Guo Y, Zhu H, Luo M, Han P, Ji X. Low-Cost Microbolometer Type Infrared Detectors. Micromachines. 2020 Sep;11(9):800. https://www.mdpi.com/2072-666X/11/9/800/pdf

Face Recognition with Thermal Imaging

A useful and quite recent study on the use of thermal imaging for facial recognition is provided in:

- Mallat K. *Efficient integration of thermal technology in facial image processing through interspectral synthesis Dissertation* (Doctoral dissertation, Sorbonne Université). https://tel.archives-ouvertes.fr/tel-03152793/document

Here we provide an exemplary receiver-operating-characteristic (ROC) curve comparing state-of-art visible FR with pure thermal and a number of inter-spectral techniques studied in this thesis.

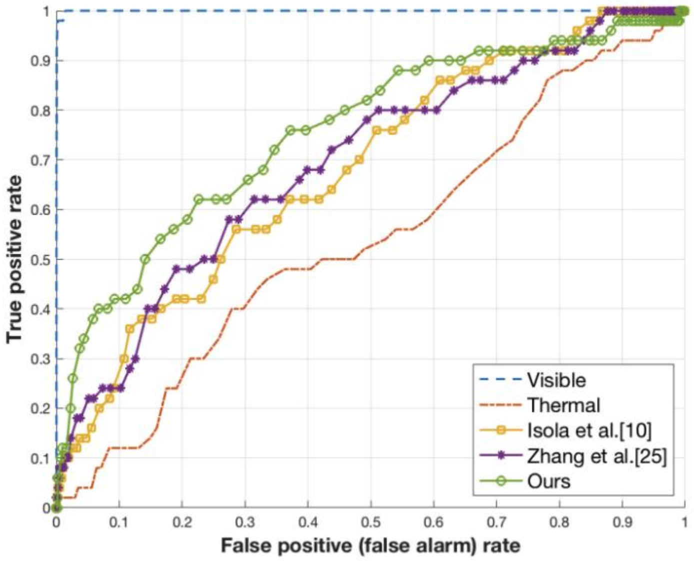


Figure 1 ROC curves for visible, thermal and combination techniques from the literature [Figure 4.7(a) taken from Mallat's thesis]

Note that the ROC curve for visible-light FR is almost ideal whereas that for pure thermal facial recognition is almost diagonal, indicating that there are nearly as many false negatives as true positives and conversely, as many false positives as true negatives for the thermal data. In essence facial recognition based on thermal image data is far less reliable than facial recognition based on visible, reflected light.

This figure and the general findings of this research work indicate that SoA for purely thermal-image based facial recognition is significantly less accurate when a like-for-like evaluation is performed, and thermal image data needs to be supplemented with additional inter-spectral data to achieve useful results. Even when supplemented in this manner it performs less effectively than FR algorithms based on visible-light and conventional digital cameras.

In the next sections some technical explanations are provided as to why thermal imaging performs so poorly for facial recognition.

Calibration of Thermal Cameras

Thermal cameras are sensitive to a wide temperature range. For uncooled microbolometers, the unconditioned sensing characteristic curve ranges from –273 degrees Celsius up to temperatures of several hundred degrees Celsius. In a thermal camera, this sensing response is adjusted to a much narrower range by adding specialized electronic circuits to condition the sensing response and calibrate the sensor output to a specific temperature range.

When a camera is adapted specifically for observing human faces it will be conditioned to a narrow temperature range around that of body temperature – typically 37 degrees +/- 5 degrees as the body temperature is not likely to rise above 40 degrees or drop below 32 degrees. With an example 10-degree temperature range, the sensitivity for 8-bit imaging is about 0.04 degree Celsius per bit-change of data for a typical thermal camera optimized for facial recognition purposes. Where there is a +/- 1-degree variation across a facial region a camera that is calibrated for FR can utilize 50 of 255 quantization levels to provide useful facial information, or about 5.5 of the 8-bit range is useful to extract relevant data.

However, the camera used to capture the data for this NUIG project was calibrated for a wider temperature range due to the use case of ambient thermal image data acquisition in roadway scenarios. Thus, the operating range is from c.–20 degrees up to +40 degrees Celsius representing a sensitivity of 0.25 degrees per bit-change of data. For this camera only c.8 of 255 quantization levels are available across a face region, or 3 of 8 bits.

It is thus clear that the level of detail that can be obtained from the thermal data recorded in the NUIG dataset, even in optimal situations, *is significantly lower than could be obtained from a typical thermal camera that is optimized for facial recognition*.

Size and Quality of Facial Crops

The facial crop size used by Mallat (Chapter 4, https://tel.archives-ouvertes.fr/tel-03152793/document) in both visible and thermal regions is 160x120 pixels (section 4.4.1 of Mallat) and this is a typical size used for facial recognition applications.

The NIST website, https://pages.nist.gov/frvt/html/frvt11.html , contains performance verifications for all public state-of-art visible-light facial recognition algorithms across a range of publicly available test datasets. NIST recommends that ideally a crop size of 240x240 pixels should be used for tests and comparisons between FR algorithms, but some of the test datasets do include faces crops of smaller size to reflect situations where a person is more distant from the camera.

The key takeaway from the NIST website is that almost all the test datasets used are > 120x120 pixels, although recently a more challenging dataset with 80x80 face crops has been added to the recommended test datasets to be used in comparison studies.

To get an additional expert opinion we have engaged in informal discussions with a local expert, Gabriel Costache from Xperi, Galway. He manages an engineering group that spent 6 months doing extensive testing of state-of-art facial recognition algorithms about 2 years ago, and continues to do bi-annual reviews on the latest advances in this field. From these discussions it was learned that their group has some level of success with reconstructing conventional facial crops as small as 45x45 pixels by employing bicubic super-resolution algorithms, and enlarging to 90x90. However, while the performance was acceptable for some use cases it was significantly impaired when compared with original 90x90 or larger face crops. These tests were based on high-quality original facial image crops with a well-focused visible-light camera which can effectively use the full 8-bits of data resolution.

Another recent study has shown that it is possible to extend visible-light FR to even smaller face crops: https://openaccess.thecvf.com/content_ICCVW_2019/papers/RLQ/Mynepalli_Recognizing_Tiny_Faces_ICCVW_2019_paper.pdf

However, the starting point for all such experiments is a set of larger high-quality visible-light face crops and the percentage of true positives for crops smaller than 30 pixels is less than 40% even with state-of-art super-resolution.

For thermal data where the dynamic range is less than 3 of 8 bits resolution (8 of 255 quantization levels for 8-bit image data) the reconstruction of any significant detail from face images smaller than 45 pixels would be completely unfeasible. For larger face regions the level of detail that could be reconstructed will be substantially lower than what could be obtained from a well calibrated thermal camera. And a well-calibrated thermal camera has a *significantly poorer performance* than that of a visible camera as discussed above.

In the next section we summarize our findings on thermal image quality and the suitability of thermal imaging for facial recognition.

Conclusions on Thermal Facial Quality

From the above discussion based on SoA research literature, the following conclusions can be drawn:

- The camera used in this study was not calibrated for Facial Recognition (FR) and thus has significantly less quantization levels available to resolve the details of facial regions than would normally be considered useful for *thermal* facial recognition.
- Given that a large facial crop is available – typically larger than 120x120 pixels – the results from Mallat show that well-callibrated thermal imaging on its own *performs very poorly* when a like-for-like study is made compared with visible-light FR.
- For visible light FR the practical limit of crop size for reconstruction to high-quality original images suitable for FR is around 30x30 pixels, but for any pragmatic levels of accuracy (i.e. > 50% true positives) at a minimum, 45x45 pixel facial crops of high quality are required.

Based on these conclusions it was decided to perform an analysis of the annotated dataset to determine any image frames with thermal facial crops that are larger than 45x45 pixels. Even though it is highly unlikely that a reconstruction of the facial details could be achieved from these face crops they will be inspected manually. Such a manual inspection will allow us to reach a determination if any of these larger regions might be frontally oriented and have a significant level of local detail on the face region.

The next section of this report details the procedures and methodology employed to review the annotated dataset, select any higher-risk face crops, and make an individual determination on each of those.

Study and Review of the Dataset

In this section the dataset is tested to determine any image frames with face crops that might be sufficiently large to pose a risk that personal face data could be reconstructed.

Methodology to Detect Face Crops > 45x45 pixels

The main thermal dataset comprises c.26,000 annotated thermal image frames of VGA (640x480) resolution. As it would not be feasible to inspect each image frame individually the first step of the review process is to apply a face-crop detector to determine which image frames have face crops and the range of crop sizes.

To detect face crops and identify their range of sizes a modified version of the MTCNN face detector was employed, and can be found at this URL:
https://github.com/JustinGuese/mtcnn-face-extraction-eyes-mouth-nose-and-speeding-it-up

The face detector is applied using the following methodology:

Step 1:	Each frame is searched using MTCNN for faces across multiple region sizes;
Step 2:	If at least one face region is detected the frame is marked positive for face crop; The detected faces are cropped and copied to a separate folder;
Step 3:	Each folder is inspected to determine the size of face crops;
Step 4:	Frames with crops > 45x45 pixels are marked and removed from the dataset;

The dataset comprises three main sub-sets, one acquired in daytime conditions, one in evening/dusk conditions and a third set in night-time conditions. The results of applying the face crop detection on each of these subsets is provided in the tables below.

Daytime Image Frames	
Total Frames	**9600**
Frames with Face Crops	**205**
Cropped Faces < 20X20	10
Cropped Faces > 20X20	26
Cropped Faces > 30x30	0
Cropped Faces > 45x45	201
Total No of Facial Crops:	237
Frames removed	201

Evening/Dusk Image Frames	
Total Frames	**11,960**
Frames with Face Crops	**526**
Cropped Faces < 20X20	320
Cropped Faces > 20X20	147
Cropped Faces > 30x30	59
Cropped Faces > 45x45	0
Total No of Facial Crops:	526
Frames removed	0

Nighttime Image Frames	
Total Frames	**4600**

Frames with Face Crops	**526**
Cropped Faces < 20X20	6
Cropped Faces > 20X20	1
Cropped Faces > 30x30	0
Cropped Faces > 45x45	0
Total No of Facial Crops:	7
Frames removed	0

It can be seen that only the first dataset has a significant number of larger face crop. These have been removed from the dataset to be made available publicly and a list of the removed image frames is provided in Appendix A.

Risk Analysis

While this report has focused on face regions within the image data there are other potential risks. Car number plates and bus numbers could be potentially identified. Each of these specific risks are discussed separately in this section. A risk assessment form is completed and appended as Appendix C.

Risk Assessment - Potential to Identify an Individual from Facial Data

This is considered the most significant risk involved in making this dataset available publicly. However the details provided in the earlier sections of this report and the measures taken to remove any larger size facial crops have protected against this risk.

As indicated previously in this report the quality of facial data recorded in thermal images acquired in this study and the level of detail that can be potentially reconstructed is so low that this risk is essentially, non-existent. Some examples are shown in Appendix B.

Risk Assessment – Reconstruction of Vehicular Number Plates

This is considered a less significant risk as, even if it were possible to partially re-construct vehicular number plates this would only identify a particular vehicle and not the driver or other occupants. But again, as thermal imaging relies on emission of thermal radiation, and as number plates are non-emissive it is unlikely that sufficient details could be discerned, even with advanced reconstruction algorithms.

On a sunny day it might be possible for dark letters and numbers to absorb heat from the sun and thus appear with a higher level of thermal emissivity than the white background of a number plate which would reflect heat form the sun. In a cloudy environment number plates and the numbers and letters are at a uniform temperature and thus no details can be discerned. As our data was acquired in overcast conditions throughout this eliminates any risks that number plate details will be discernible.

Separately it is challenging even with conventional digital-cameras, to discern and reconstruct number plates in normal traffic conditions from lower resolution cameras, such as the VGA (640x480 pixels) resolution available from the thermal imager. An example comparison between conventional camera imaging of number plates, compared with thermal imaging is provided in Appendix B.

Our conclusion is that both the risk here and the resulting consequence are extremely low.

Risk Assessment – Reconstruction of Public Transport Bus Numbers

It may be feasible to determine the route numbers of public bus transportation due to the use of LED illumination for these numbers as the lighting may be thermally emissive. However from our inspection of several examples in the dataset this seems to be unlikely as the glass cover for these numbers on the bus appears to absorb any thermal emissivity. A comparative example is provided in Appendix B showing a very faint level of detail.

Even if it were feasible to reconstruct bus numbers this only reveals the bus-route, but not any passenger identities or even the bus driver identity as no time or date information is provided with this dataset. Thus given the knowledge that the data was gathered in Galway town, and even if bus routes could be identified, the associated consequences of this aren't of significance in terms of personal data privacy.

Appendix A – List of Frames Removed from the Dataset

Daytime Image Frames	Frame Number
Sequence 1	from-car-1 2410
	from-car-1 2411
	from-car-1 2412
	from-car-1 2413
	from-car-1 2414
	from-car-1 2415
	from-car-1 2416
	from-car-1 2417
	from-car-1 2418
	from-car-1 2419
	from-car-1 2420
	from-car-1 2421
	from-car-1 2422
	from-car-1 2423
	from-car-1 2424
	from-car-1 2425
	from-car-1 2425
Sequence 2	from-car-1-4813
	from-car-1-4814
	from-car-1-4815
	from-car-1-4816
	from-car-1-4817
	from-car-1-4818
	from-car-1-4819
	from-car-1-4820
	from-car-1-4821
	from-car-1-4822
	from-car-1-4823
	from-car-1-4824
	from-car-1-4825
	from-car-1-4826
	from-car-1-4827
	from-car-1-4828
	from-car-1-4829
	from-car-1-4830
	from-car-1-4831
	from-car-1-4832
	from-car-1-4833
	from-car-1-4834
	from-car-1-4835
	from-car-1-4836
	from-car-1-4837
	from-car-1-4838
	from-car-1-4839
	from-car-1-4840
	from-car-1-4841
	from-car-1-4842
	from-car-1-4843
	from-car-1-4844
	from-car-1-4845
	from-car-1-4846
	from-car-1-4847
	from-car-1-4848
	from-car-1-4849
	from-car-1-4850
	from-car-1-4851
	from-car-1-4852
	from-car-1-4853

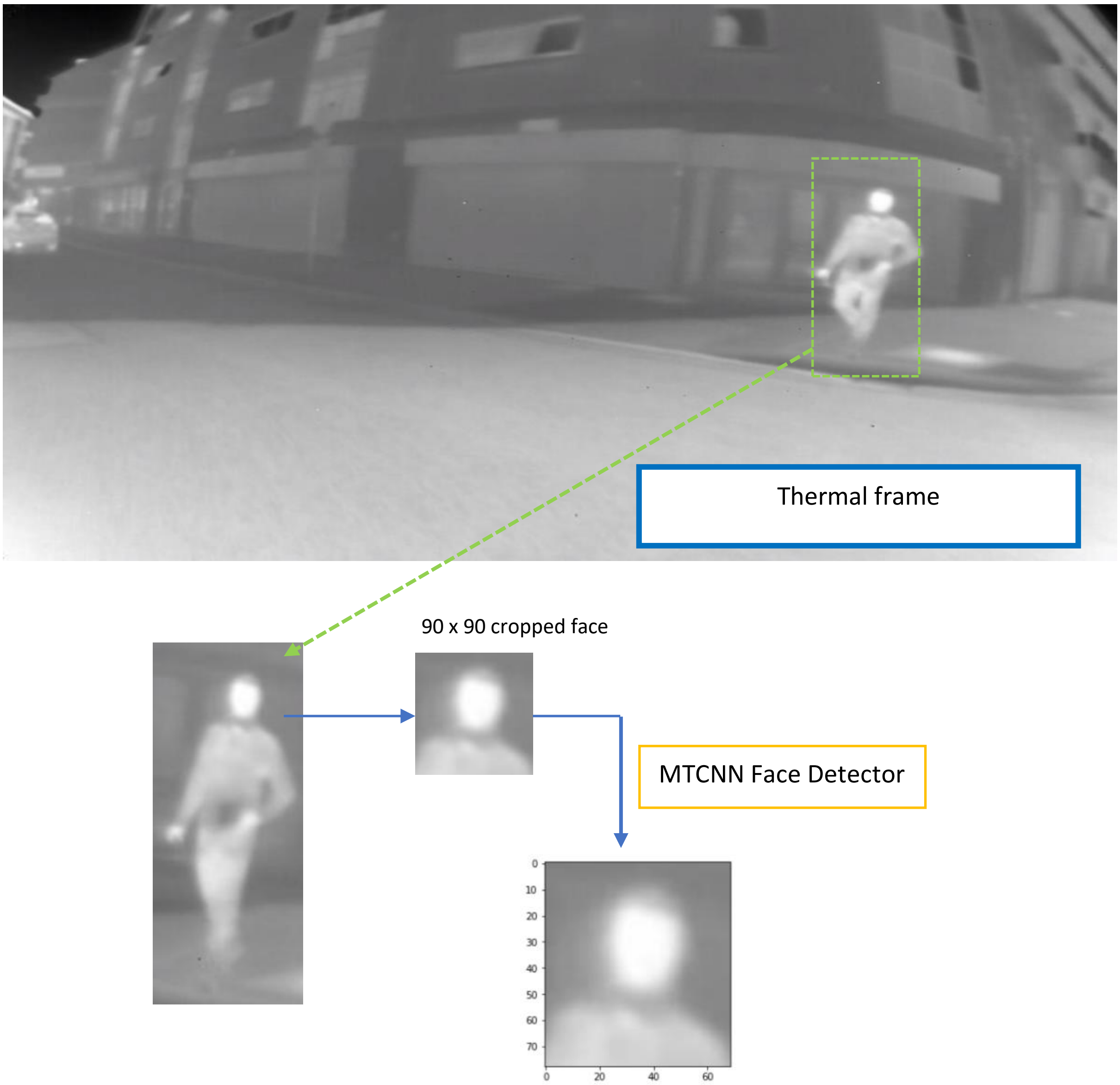
	from-car-1-4854
	from-car-1-4855
	from-car-1-4856
	from-car-1-4857
	from-car-1-4858
	from-car-1-4859
	from-car-1-4860
	from-car-1-4861
	from-car-1-4862
	from-car-1-4863
	from-car-1-4864
	from-car-1-4865
	from-car-1-4866
	from-car-1-4867
	from-car-1-4868
	from-car-1-4869
	from-car-1-4870
	from-car-1-4871
	from-car-1-4872
	from-car-1-4873
	from-car-1-4874
	from-car-1-4875
	from-car-1-4876
	from-car-1-4877
	from-car-1-4878
	from-car-1-4879
	from-car-1-4880
	from-car-1-4881
	from-car-1-4882
	from-car-1-4883
	from-car-1-4884
	from-car-1-4885
	from-car-1-4886
	from-car-1-4887
	from-car-1-4888
	from-car-1-4889
	from-car-1-4890
	from-car-1-4891
	from-car-1-4892
	from-car-1-4893
	from-car-1-4894
	from-car-1-4894
	from-car-1-4896
	from-car-1-4897
	from-car-1-4898
	from-car-1-4899
	from-car-1-4898
	from-car-1-4900
	from-car-1-4901
	from-car-1-4902
	from-car-1-4903
	from-car-1-4904
	from-car-1-4905
	from-car-1-4899
	from-car-1-4898
	from-car-1-4900
	from-car-1-4901
	from-car-1-4902
	from-car-1-4903
	from-car-1-4904
	from-car-1-4905
	from-car-1-4899
	from-car-1-4898
	from-car-1-4900

	from-car-1-4901
	from-car-1-4902
	from-car-1-4903
	from-car-1-4904
	from-car-1-4905
	from-car-1-4906
	from-car-1-4907
	from-car-1-4908
	from-car-1-4909
	from-car-1-4910
	from-car-1-4911
	from-car-1-4912
	from-car-1-4913
	from-car-1-4914
	from-car-1-4915
	from-car-1-4916
	from-car-1-4917
	from-car-1-4918
	from-car-1-4919
	from-car-1-4920
	from-car-1-4921
	from-car-1-4922
	from-car-1-4923
	from-car-1-4924
	from-car-1-4925
	from-car-1-4926
	from-car-1-4927
	from-car-1-4928
	from-car-1-4929
	from-car-1-4930
	from-car-1-4931
	from-car-1-4932
	from-car-1-4933
	from-car-1-4934
	from-car-1-4935
	from-car-1-4936
	from-car-1-4937
	from-car-1-4938
	from-car-1-4939
	from-car-1-4940
	from-car-1-4941
	from-car-1-4942
	from-car-1-4943
	from-car-1-4944
	from-car-1-4945
	from-car-1-4946
	from-car-1-4947
	from-car-1-4948
	from-car-1-4949
	from-car-1-4950
	from-car-1-4951
	from-car-1-4952
	from-car-1-4953
	from-car-1-4954
	from-car-1-4955
	from-car-1-4956
	from-car-1-4957
	from-car-1-4958
	from-car-1-4959
	from-car-1-4960
	from-car-1-4961
	from-car-1-4962
	from-car-1-4963
	from-car-1-4964

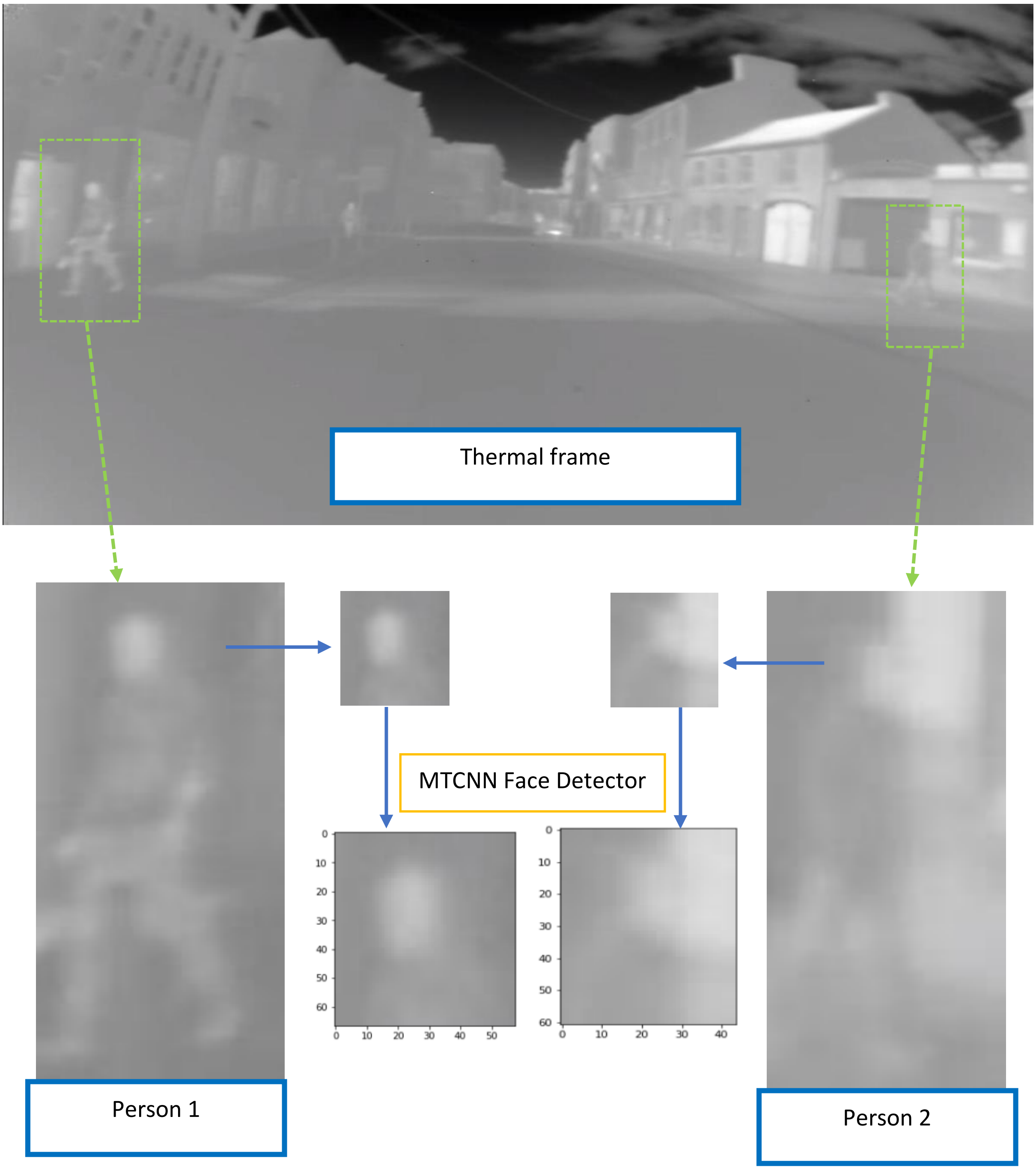
	from-car-1-4965
	from-car-1-4966
	from-car-1-4967
	from-car-1-4968
	from-car-1-4969
	from-car-1-4970
	from-car-1-4971
	from-car-1-4972
	from-car-1-4973
	from-car-1-4974
	from-car-1-4975
	from-car-1-4976
	from-car-1-4977
	from-car-1-4978
	from-car-1-4979
	from-car-1-4980
	from-car-1-4981
	from-car-1-4982
	from-car-1-4983
	from-car-1-4984
	from-car-1-4985
	from-car-1-4986
	from-car-1-4987
	from-car-1-4988
	from-car-1-4989
	from-car-1-4990
	from-car-1-4991
	from-car-1-4992
	from-car-1-4993
	from-car-1-4994
	from-car-1-4995
	from-car-1-4996

Appendix B – Data Examples

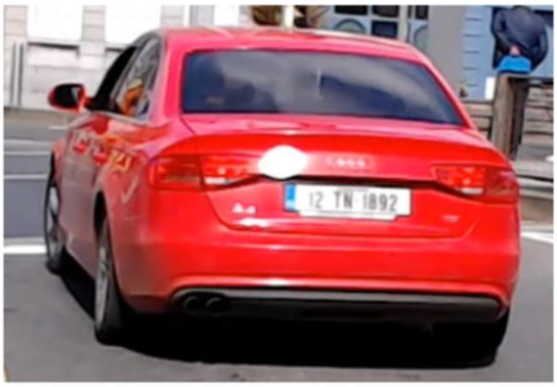
Thermal Face Crops – Example #1



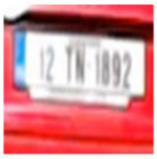
Thermal Face Crops – Example #2



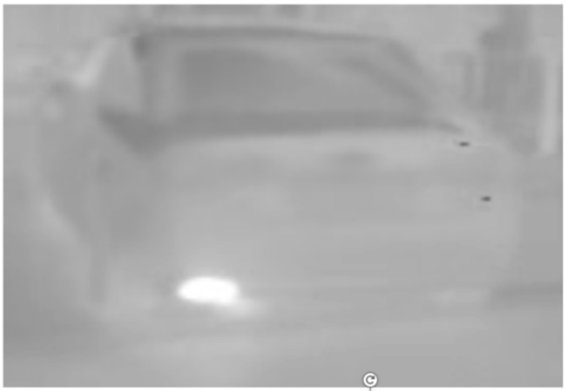
RGB Vs Thermal Image Comparison - Number Plate Data



Visible Frame



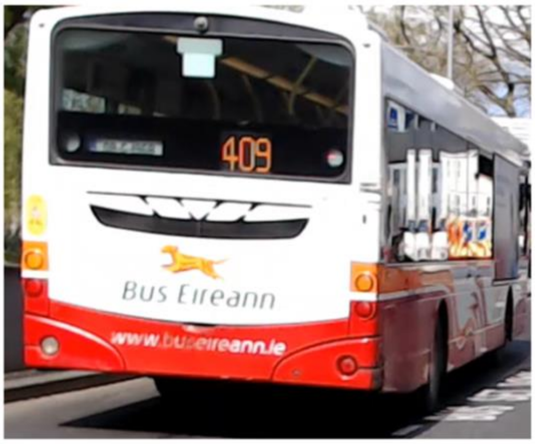
Car number plate cropped 90 x 90 visible frame

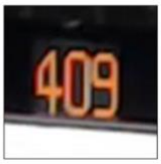


Thermal frame

Car number plate cropped 90 x 90 thermal frame

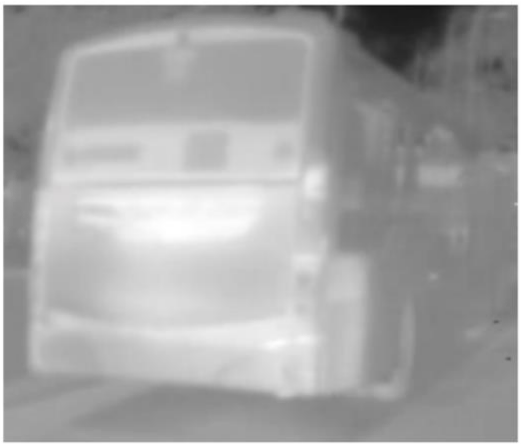
RGB Vs Thermal Image Comparison – Bus Number Detail





Bus number cropped 90 x 90 visible frame

Visible frame



Bus number cropped 90 x 90 thermal frame

Thermal frame

Appendix C – Risk Assessment

HAZARD	HAZARD OUTCOME	RISK ASSESSMENT CRITERIA A. Likelihood B. Severity/ Consequence of exposure		RISK ASSESSMENT	CONTROLS/ARRANGEMENTS	PERSON RESPONSIBLE	RESOURCES
		A	B	(A X B)	□ List existing controls according to the "hierarchy of controls". □ Indicate status of measures To do - action date given Completed - √ Ongoing - underlined		
Reconstruction of facial data to allow identification of individual persons	Individual persons recognized triggering GDPR violation	< 1	4	< 4	Considered a minor risk. From an abundance of caution, remove all face regions > 45x45 pixels (done)		
Number plate reconstruction	Individual vehicle identified with indirect association to vehicle owner	< 1	3	< 3	Considered a minor risk as no explicit GDPR Violation (not possible to determine driver).		
Public bus number reconstruction	Vehicle identified	<1	1	<1	Considered trivial risk; no explicit GDPR Violation (not possible to identify passengers).		
Building Identification	Person identified by proximity to building	<1	1	<1	Considered trivial risk; no explicit GDPR Violation; no time or date information so not possible to definitively associate a person with a particular building.		

Likelihood	Guide Description
5	Very likely/imminent – certain to happen
4	Probable – a strong possibility of it happening
3	Possible – it may have happened before
2	Unlikely - could happen but unusual
1	Rare – highly unlikely to occur

Severity	Guide Description
5	Catastrophic - fatality, catastrophic damage
4	Major – significant injury or property damage, hospitalisation
3	Moderate - injury requiring further treatment, lost time
2	Minor - first aid injury, no lost time
1	Very minor – insignificant injury

		Severity (S)				
		1	**2**	**3**	**4**	**5**
Likelihood (L)	**5**	5	10	15	20	25
	4	4	8	12	16	20
	3	3	6	9	12	15
	2	2	4	6	8	10
	1	1	2	3	4	5

Risk Rating (RR)	**Action**
High Risk	Stop the task/activity until controls can be put into place to reduce the risk to an acceptable level
Medium Risk	Determine if further safety precautions are required to reduce risk to as low as is reasonably practicable
Low Risk	No further action, keep under review

Signature of Risk Assessor	Peter Corcoran	Name / job title:	Director C3Imaging Group, NUIG
Details of any persons consulted	Gabriel Costache, Director Engineering, Xperi Ltd., Galway		

Appendix K

Subject Consent Forms

Farooq, Muhammad Ali

From:	Adrian Ungureanu <ung.adrian@yahoo.com>
Sent:	Wednesday 19 February 2020 14:35
To:	Farooq, Muhammad Ali
Subject:	Re: Thermal Facial Database Permission Required for GDPR

Hello, Ali

I confirm that you can use those pictures of me in your work.

I confirm that the nature, demand and possible risks of the research have been explained to me and I understand and accept them. I understand that my consent is entirely voluntary and that I may withdraw at any time from the research project without proper explanation or penalty. I also consent that my data will be publicly available for further research studies

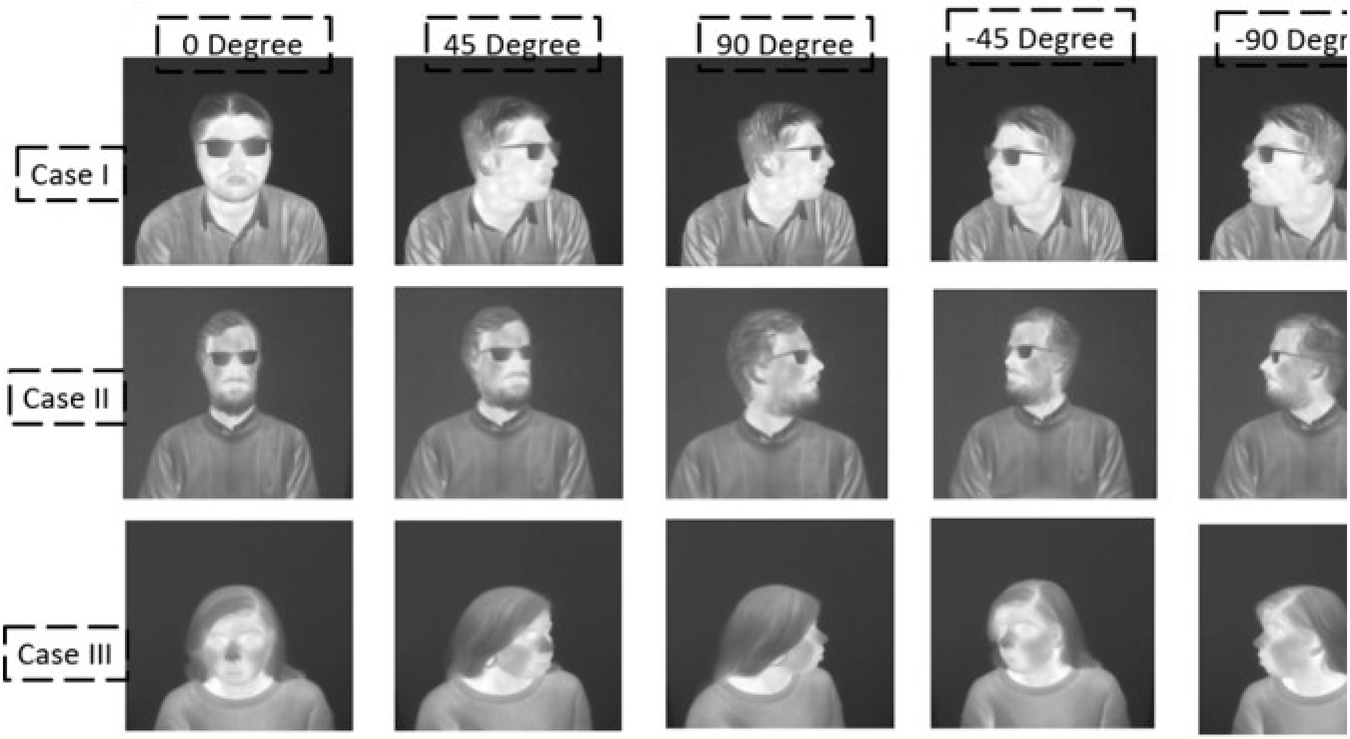
All the best,
Adrian

Sent from Yahoo Mail on Android

On Wed, Feb 19, 2020 at 14:05, Farooq, Muhammad Ali
<M.Farooq3@nuigalway.ie> wrote:

> Hello Friends,
>
> I hope you all are doing well.
>
> Currently, I am finalizing my publication under Heliaus Project titled as
>
> "Performance Estimation of the State of Art Convolution Neural Networks (CNN) for Thermal Images-Based Gender Classification System"
>
> I have gathered your datasets in the lab and I am using it my paper.

These are the facial poses from different angles as shown in the below figure



Of course, it does not includes visible images (Just thermal).

If you guys are willing to get these images published it and available on GitHub repository you just have to reply this email with the following lines"

" I confirm that I have read the information provided dated _______________ for the above study. I confirm that the nature, demand and possible risks of the research have been explained to me and I understand and accept them. I understand that my consent is entirely voluntary and that I may withdraw at any time from the research project without proper explanation or penalty. I also consent that my data will be publicly available for further research studies"

Thank you

Best Regards,

Muhammad Ali Farooq

PhD Researcher

HELIAUS Project

College of Engineering and Informatics

National University of Ireland Galway (NUIG)

Farooq, Muhammad Ali

From:	KHATOON, ASMA
Sent:	Wednesday 19 February 2020 15:21
To:	Farooq, Muhammad Ali; Adrian Ungureanu
Cc:	Corcoran, Peter
Subject:	Re: Thermal Facial Database Permission Required for GDPR

Hi Ali,

Please see my response below. I don't exactly remember the date when you collected the data so you can put the date yourself.

" I confirm that I have read the information provided dated ________________ for the above study. I confirm that the nature, demand and possible risks of the research have been explained to me and I understand and accept them. I understand that my consent is entirely voluntary and that I may withdraw at any time from the research project without proper explanation or penalty. I also consent that my data will be publicly available for further research studies"

Best Regards,
Asma

From: Farooq, Muhammad Ali <M.Farooq3@nuigalway.ie>
Sent: Wednesday, February 19, 2020 2:05:09 PM
To: Adrian Ungureanu <ung.adrian@yahoo.com>; KHATOON, ASMA <A.KHATOON1@nuigalway.ie>
Cc: Corcoran, Peter <peter.corcoran@nuigalway.ie>
Subject: Thermal Facial Database Permission Required for GDPR

Hello Friends,

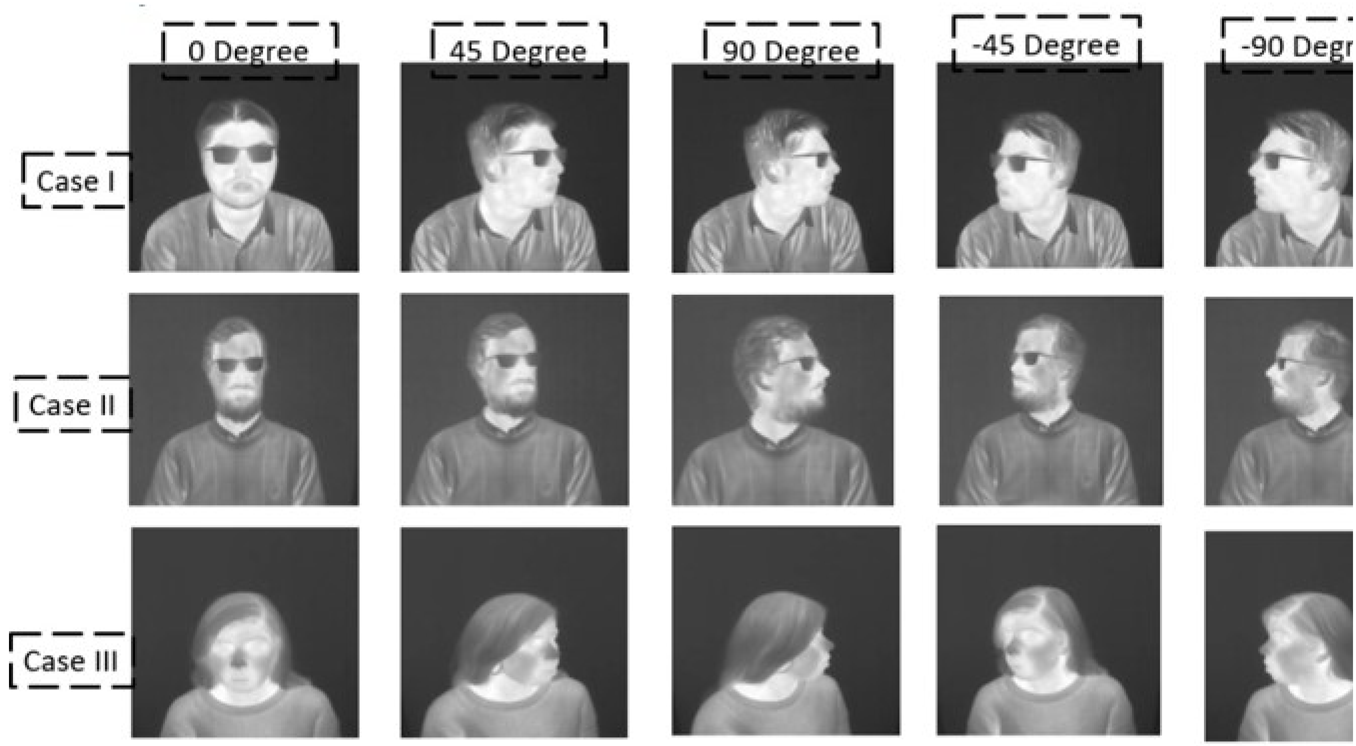
I hope you all are doing well.

Currently, I am finalizing my publication under Heliaus Project titled as

"Performance Estimation of the State of Art Convolution Neural Networks (CNN) for Thermal Images-Based Gender Classification System"

I have gathered your datasets in the lab and I am using it my paper.

These are the facial poses from different angles as shown in the below figure



Of course, it does not includes visible images (Just thermal).

If you guys are willing to get these images published it and available on GitHub repository you just have to reply this email with the following lines”

“ I confirm that I have read the information provided dated ________________ for the above study. I confirm that the nature, demand and possible risks of the research have been explained to me and I understand and accept them. I understand that my consent is entirely voluntary and that I may withdraw at any time from the research project without proper explanation or penalty. I also consent that my data will be publicly available for further research studies”

Thank you

Best Regards,
Muhammad Ali Farooq
PhD Researcher
HELIAUS Project
College of Engineering and Informatics
National University of Ireland Galway (NUIG)

Farooq, Muhammad Ali

From:	Andrade, Evismar
Sent:	Monday 24 February 2020 12:35
To:	Farooq, Muhammad Ali
Subject:	Re: Thermal Facial Database Permission Required for GDPR

I confirm that I have read the information provided dated 24/02/2020 for the above study. I confirm that the nature, demand and possible risks of the research have been explained to me and I understand and accept them. I understand that my consent is entirely voluntary and that I may withdraw at any time from the research project without proper explanation or penalty. I also consent that my data will be publicly available for further research studies"

Regards,
Evismar

From: "Farooq, Muhammad Ali" <M.Farooq3@nuigalway.ie>
Date: Monday 24 February 2020 at 12:32
To: "Andrade, Evismar" <evismar.andrade@nuigalway.ie>, "ANDRADE, EVISMAR" <E.ANDRADE1@nuigalway.ie>
Cc: Muhammad Ali Farooq <MuhammadAli.Farooq@xperi.com>
Subject: FW: Thermal Facial Database Permission Required for GDPR

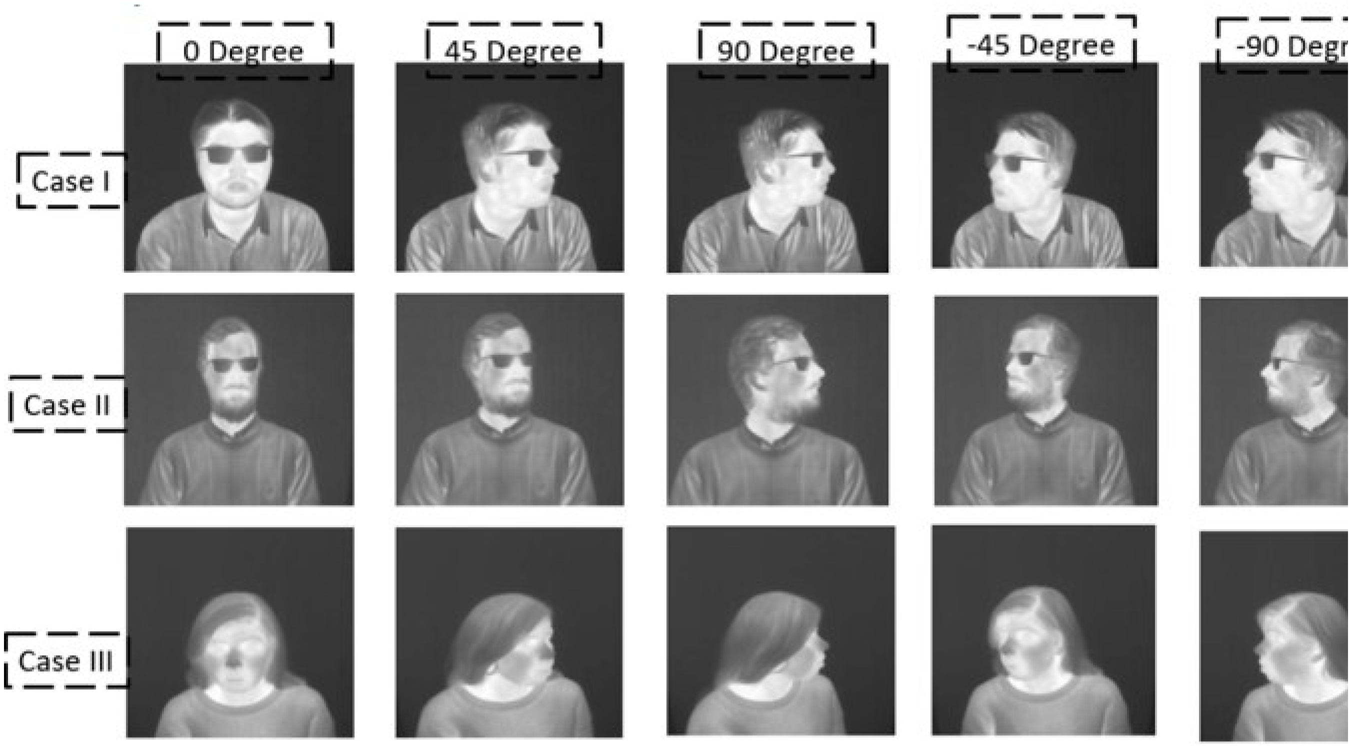
Hello Evismar,

I hope this email finds you well.

Currently, I am finalizing my publication under Heliaus Project titled as

"Performance Estimation of the State of Art Convolution Neural Networks (CNN) for Thermal Images-Based Gender Classification System"

These are the facial poses from different angles as shown in the below figure



Of course, it does not includes visible images (Just thermal).

If you are willing to get these images published it and available on GitHub repository you just have to reply this email with the following lines"

" I confirm that I have read the information provided dated ________________ for the above study. I confirm that the nature, demand and possible risks of the research have been explained to me and I understand and accept them. I understand that my consent is entirely voluntary and that I may withdraw at any time from the research project without proper explanation or penalty. I also consent that my data will be publicly available for further research studies"

Thank you

Best Regards,
Muhammad Ali Farooq
PhD Researcher
HELIAUS Project
College of Engineering and Informatics
National University of Ireland Galway (NUIG)

Farooq, Muhammad Ali

From:	Moustafa, Mohamed
Sent:	Monday 18 October 2021 13:58
To:	Farooq, Muhammad Ali
Subject:	Re: Thermal Facial Database Permission Required for GDPR

I Mohamed Moustafa confirm that I have read the information provided dated 18/10/2021 for the above study. I confirm that the nature, demand, and possible risks of the research have been explained to me and I understand and accept them. I understand that my consent is entirely voluntary and that I may withdraw at any time from the research project without proper explanation or penalty. I also consent that my data will be publicly available for further research studies

From: Farooq, Muhammad Ali <M.Farooq3@nuigalway.ie>
Sent: Monday, October 18, 2021 1:54:20 PM
To: Moustafa, Mohamed <M.Moustafa1@nuigalway.ie>
Subject: Thermal Facial Database Permission Required for GDPR

Hello Mohammad,

I hope you all are doing well.

Currently, I am in the writing phase of my thesis report. During my experimental work, I have collected your thermal facial data for various experiments. Now I need your approval for using your thermal facial pictures.

Of course, it does not includes visible images (Just thermal).

If you are willing to get these images published in my Ph.D. report you just have to reply to this email with the following lines"

" I ___________ confirm that I have read the information provided dated ________________ for the above study. I confirm that the nature, demand, and possible risks of the research have been explained to me and I understand and accept them. I understand that my consent is entirely voluntary and that I may withdraw at any time from the research project without proper explanation or penalty. I also consent that my data will be publicly available for further research studies"

Thank you

Best Regards,
Muhammad Ali Farooq
PhD Researcher
HELIAUS Project
College of Engineering and Informatics

Farooq, Muhammad Ali

From:	Paul Kielty <Paul.Kielty@xperi.com>
Sent:	Wednesday 19 February 2020 17:11
To:	Farooq, Muhammad Ali
Subject:	Re: Thermal Facial Database Permission Required for GDPR

I confirm that I have read the information provided dated 19/02/2020 for the above study. I confirm that the nature, demand and possible risks of the research have been explained to me and I understand and accept them. I understand that my consent is entirely voluntary and that I may withdraw at any time from the research project without proper explanation or penalty. I also consent that my data will be publicly available for further research studies.

Paul Kielty

From: Farooq, Muhammad Ali <M.Farooq3@nuigalway.ie>
Sent: Wednesday 19 February 2020 16:52
To: Paul Kielty <Paul.Kielty@xperi.com>
Subject: Thermal Facial Database Permission Required for GDPR

This message has originated from an **External Source**. Please use proper judgment and caution when opening attachments, clicking links, or responding to this email.

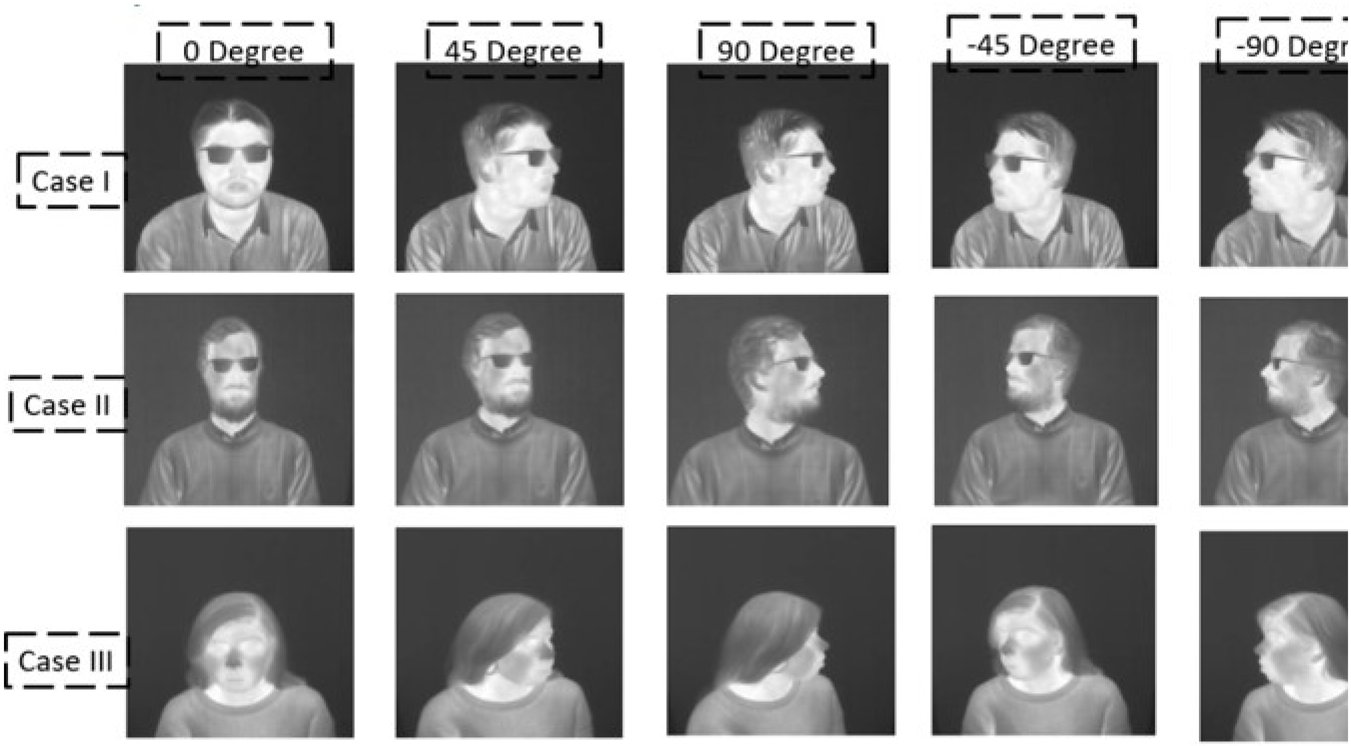
Hello Friends,

I hope you all are doing well.

Currently, I am finalizing my publication under Heliaus Project titled as

"Performance Estimation of the State of Art Convolution Neural Networks (CNN) for Thermal Images-Based Gender Classification System"

These are the facial poses from different angles as shown in the below figure



Of course, it does not includes visible images (Just thermal).

If you are willing to get these images published it and available on GitHub repository you just have to reply this email with the following lines"

" I confirm that I have read the information provided dated ________________ for the above study. I confirm that the nature, demand and possible risks of the research have been explained to me and I understand and accept them. I understand that my consent is entirely voluntary and that I may withdraw at any time from the research project without proper explanation or penalty. I also consent that my data will be publicly available for further research studies"

Thank you

Best Regards,
Muhammad Ali Farooq
PhD Researcher
HELIAUS Project
College of Engineering and Informatics
National University of Ireland Galway (NUIG)

Appendix L

Additional Experimental Work Related to Face Localization and Facial Landmark Detection in Thermal Images

Face Localization and Facial Landmark Detection on Thermal Data for In-Cabin Driver Monitoring Application

This study presents the working methodology and experimental results of effectual face detection and facial landmarks detection algorithms adapted for thermal facial data for in-cabin driver monitoring applications related to WP-7 of the Heliaus project. The efficacy of these algorithms is validated on the thermal data acquired from the prototype LWIR uncooled thermal camera.

1. *Working Methodology*

This section will be explaining the working methodology of non-contact thermal facial detection systems. Figure 1 shows the complete workflow diagram of the proposed system.

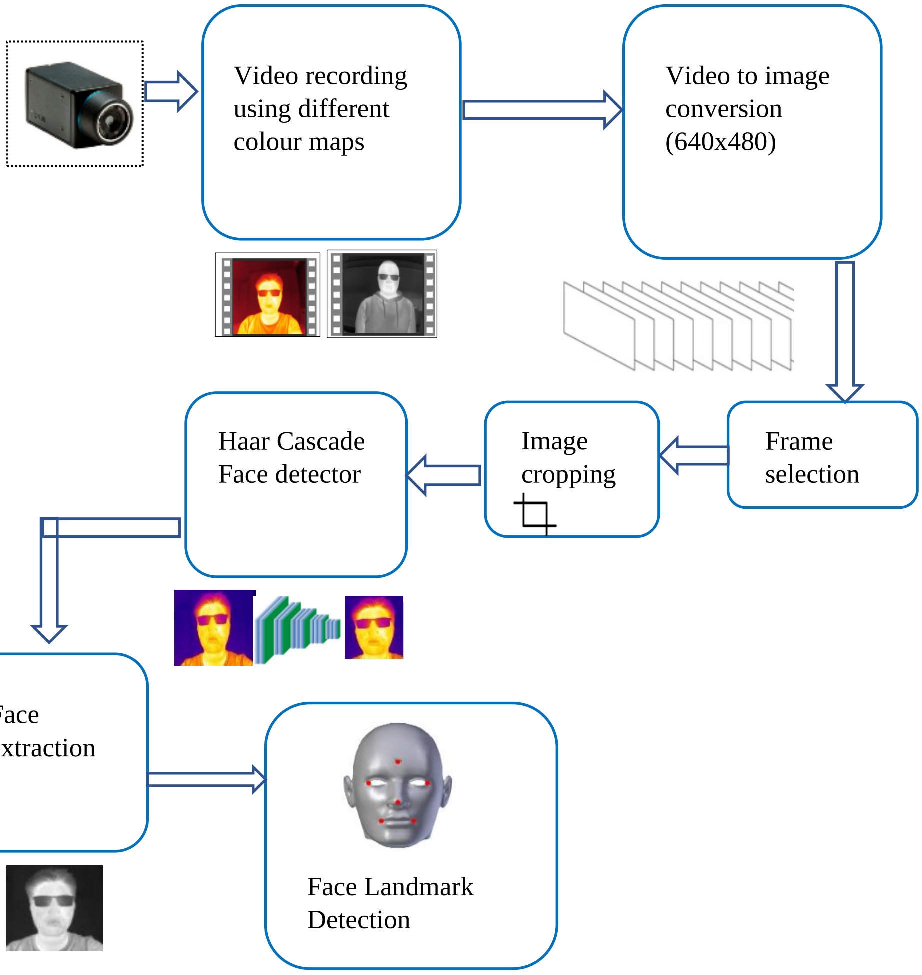

Video recording using different colour maps
Video to image conversion (640x480)
Frame selection
Image cropping
Haar Cascade Face detector
Face extraction
Face Landmark Detection

Figure 1: Block diagram representation of face localization and facial landmarks detection on thermal data.

The acquired video sets from the prototype thermal camera (as explained in section 3.3.1 of chapter 3) are first converted to image data. Once the frames are created, the next phase includes selecting the best set of thermal frames showing different facial angles. The third step includes image cropping which is done using the OpenCV library. It is done to select the upper body and mainly the facial area (as the region of interest). The frames are cropped into 400x300 resolution. The next step includes applying the HAAR Cascade [65] to detect and localize the face area in the thermal frame. Lastly, we have applied an end-to-end CNN network to detect the facial landmarks on the extracted face frame using the Haar Cascade face detector.

2. Face Detector

This section will explain the working principle of the HAAR Cascade face detector algorithm to perform face detection in the desired frame. It is achieved by employing a haar feature-based cascade classifier. The algorithm was proposed by [65] and is based on the Viola-Jones detection method [65]. It is based on the Haar wavelet technique to analyse pixels in the image into squares by function. It uses important image concepts to compute the features detected. Haar Cascades uses the Ada-boost learning algorithm which selects a small number of important features from a large training set to give an efficient result. In the second stage, it uses cascading techniques to detect the face in an image. It can be trained and used for various computer vision applications such as palm detection, and object detection tasks. The network is originally trained on RGB data, however, in this study, we have used the viola jones algorithm to validate its effectiveness on thermal facial data.

3. Face Localization and Facial Landmarks Detection Results

This section will present the result analysis of face detector algorithms on thermal frames of different subjects acquired from prototype uncooled LWIR thermal imaging sensor. Once the data is acquired, a subset of 70 frames was selected which was extracted from 6 different video recordings where the haar cascade face detector was applied. The results demonstrate the optimal performance in the form of precise face detection/ localization results. For this purpose, a special Graphical User Interface (GUI) was designed in MATLAB R2018 which was installed on a Core I7 machine with 32GB Ram. The system GUI is depicted in Figure 2.

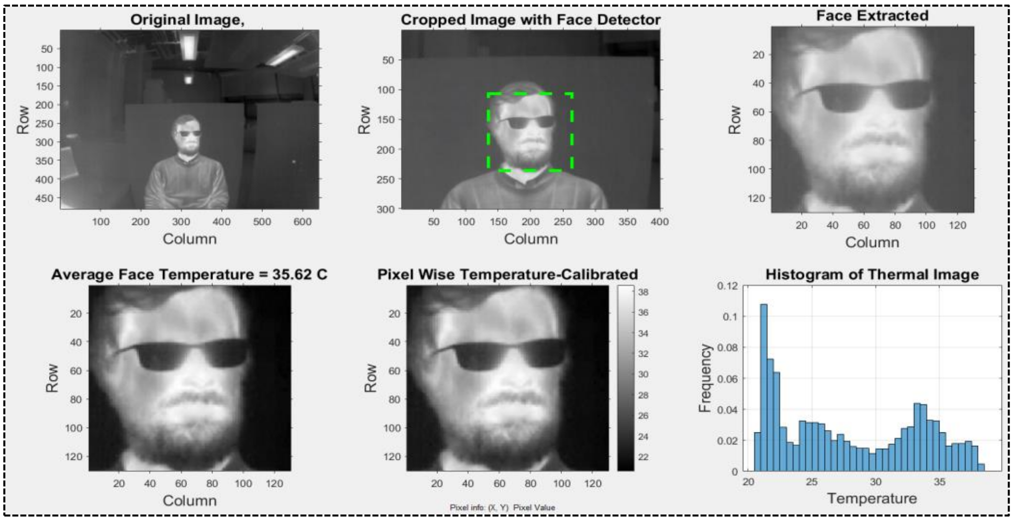
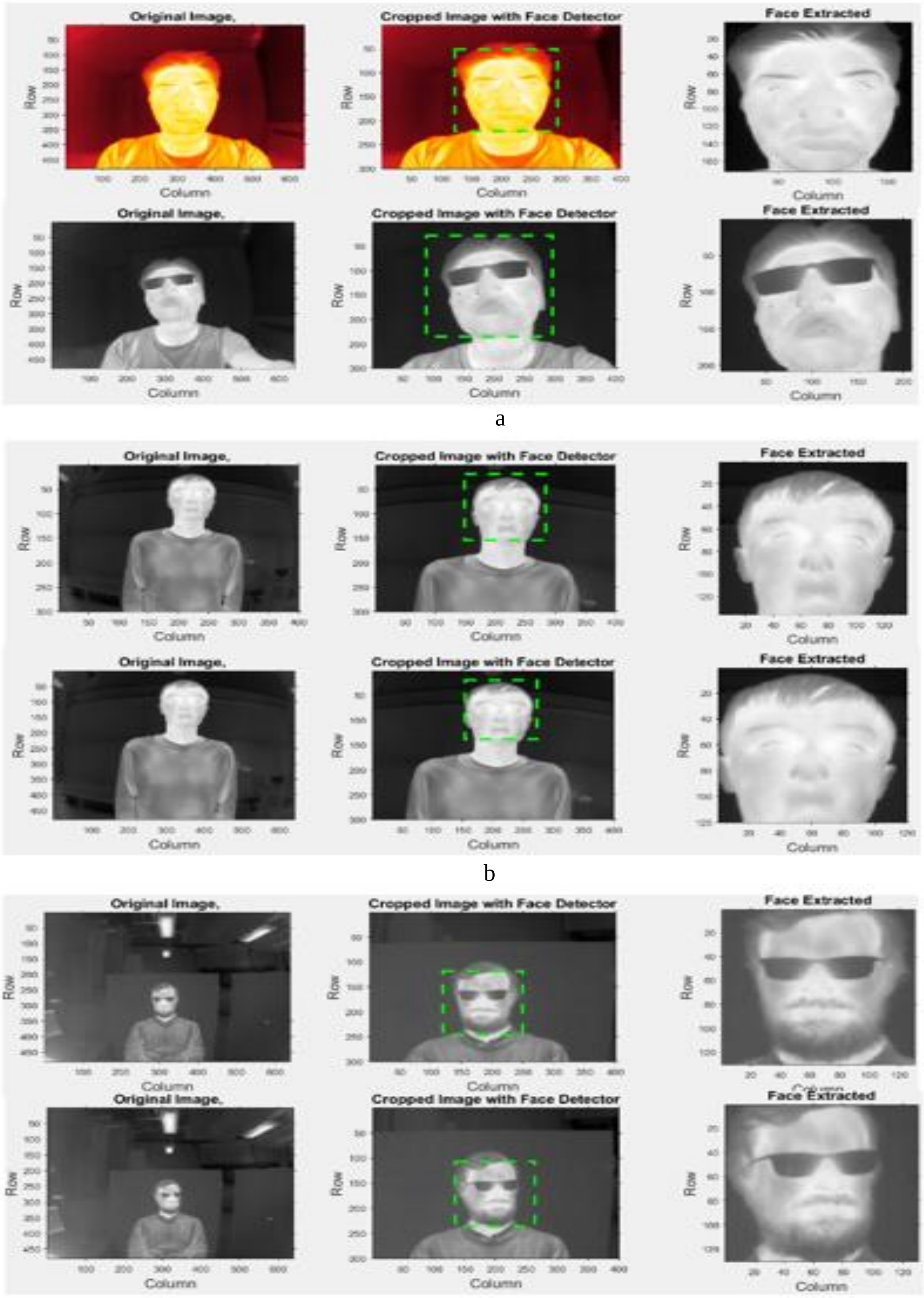


Figure 2: Complete system Graphical User Interface (GUI).

Figure 3 shows the face detection results on various face poses of three different male subjects using the Haar Cascade Face detector. The results were acquired using Matlab software.



a

b

c

Figure 3: Haar Cascade face detector results on three different male subjects, (a) subject-1 at 60 cm distance with two different facial poses and with and without glasses, (b) Subject-2 at 70cm distance with different frames and without glasses, (c) Subject-3 at 80cm distance with different facial poses and with glasses.

Figure 4 shows the histogram representation of the overall temperature distribution of the cropped thermal face region using the HAAR Cascade face detector algorithm. The results show the histogram representation on the cropped region of interest of two male subjects captured from the LWIR uncooled prototype thermal camera in an indoor lab environment as demonstrated in Figure 4.

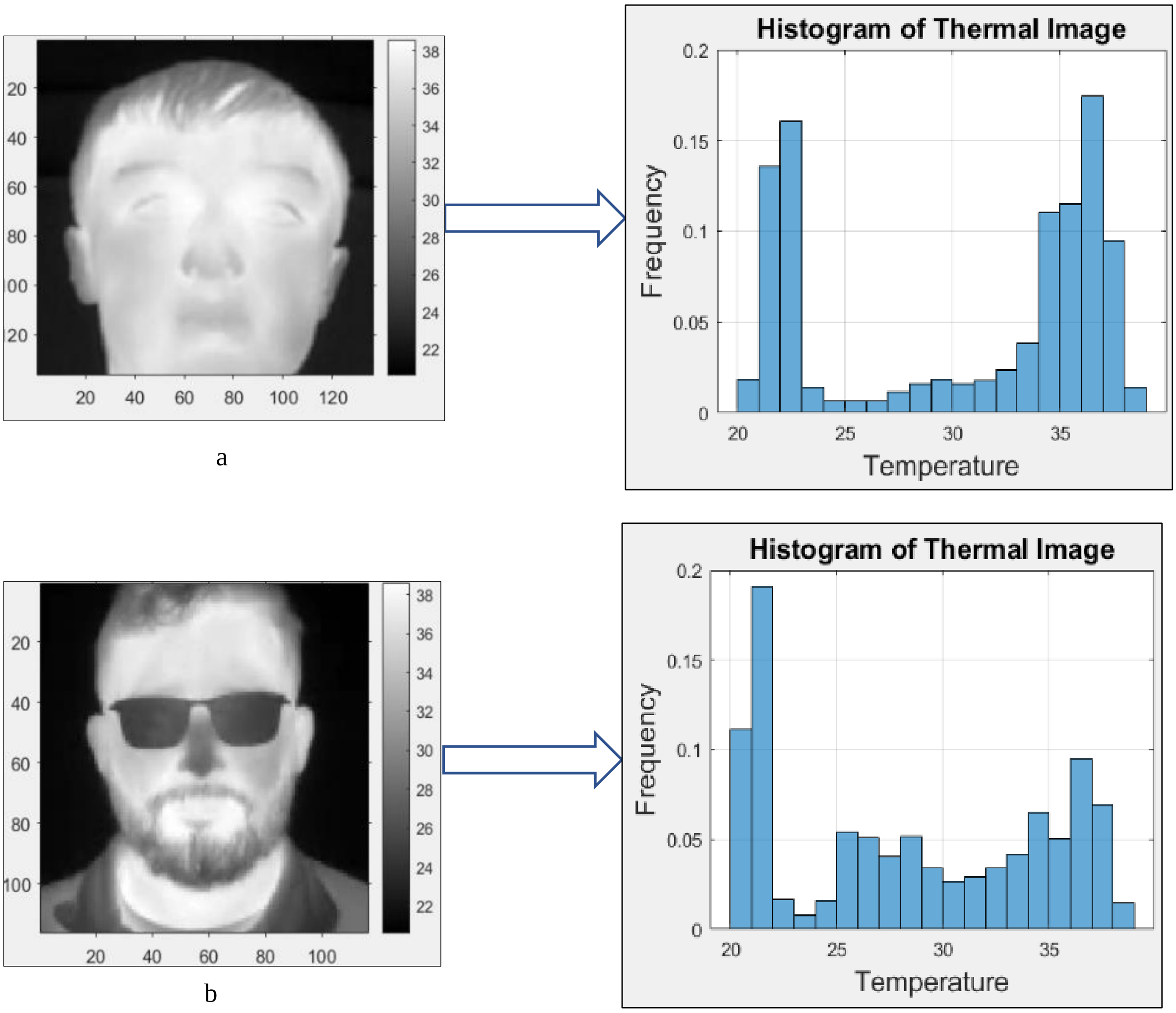


Figure 4: Facial temperature distribution histogram (a) facial thermal image of male subject A with minimum/ maximum temperature scale and Histogram of thermal images showing temperature distribution in the form of grey-level distribution, (b) facial thermal image of male subject B with respective temperature distribution histogram of the cropped facial region.

The second phase of the experimental results shows the key facial features by detecting 68 facial landmarks using the localized thermal facial region. The facial landmark detector is applied using the dlib library which is inspired by the published study of Kazemi and Sullivan (2014) titled 'One Millisecond Face Alignment with an Ensemble of Regression Trees' [66]. This method works in two steps which are as follows.

1. A training set of labeled facial landmarks on an image. These images are manually labeled, specifying specific (x, y)-coordinates of regions surrounding each facial shape.
2. Given this training data, an ensemble of regression trees are trained to estimate the facial landmark positions directly from the pixel intensities themselves (i.e., no "feature extraction" is taking place in this method).

The output results of this method produce facial landmarks either offline or in real-time with high-quality predictions. In this research work, we had used the pretrained facial landmark detector by using the dlib library to estimate the location of 68 (x, y)-coordinates that map to facial structures to validate its efficacy on thermal data. Figure 5 shows the six facial landmarks located at six different face key features.

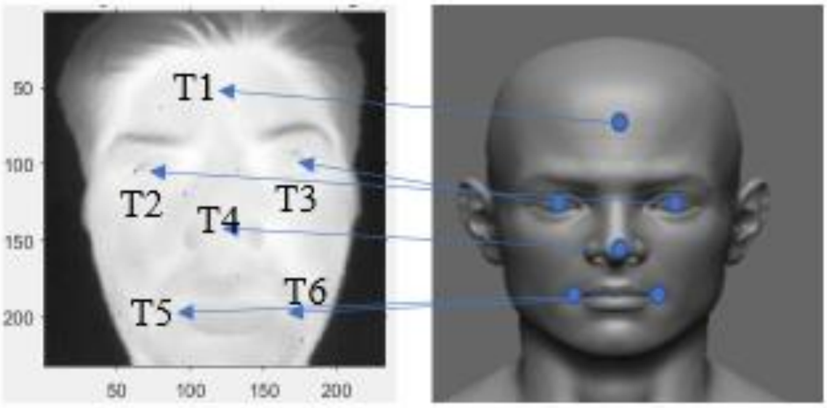


Figure 5: Six different facial points marked as T1-T6, 1) Forehead, 2) center of the right eye, 3) center of the left eye, 4) tip of the nose, 5) right side of the lip, 6) left side of the lips.

Figure 6 shows the results of 68 (x, y)-coordinates facial landmark detector on six thermal frames extracted from an acquired video sequence of a male subject wearing glasses. These thermal frames show different facial angles and varying distances from the camera.

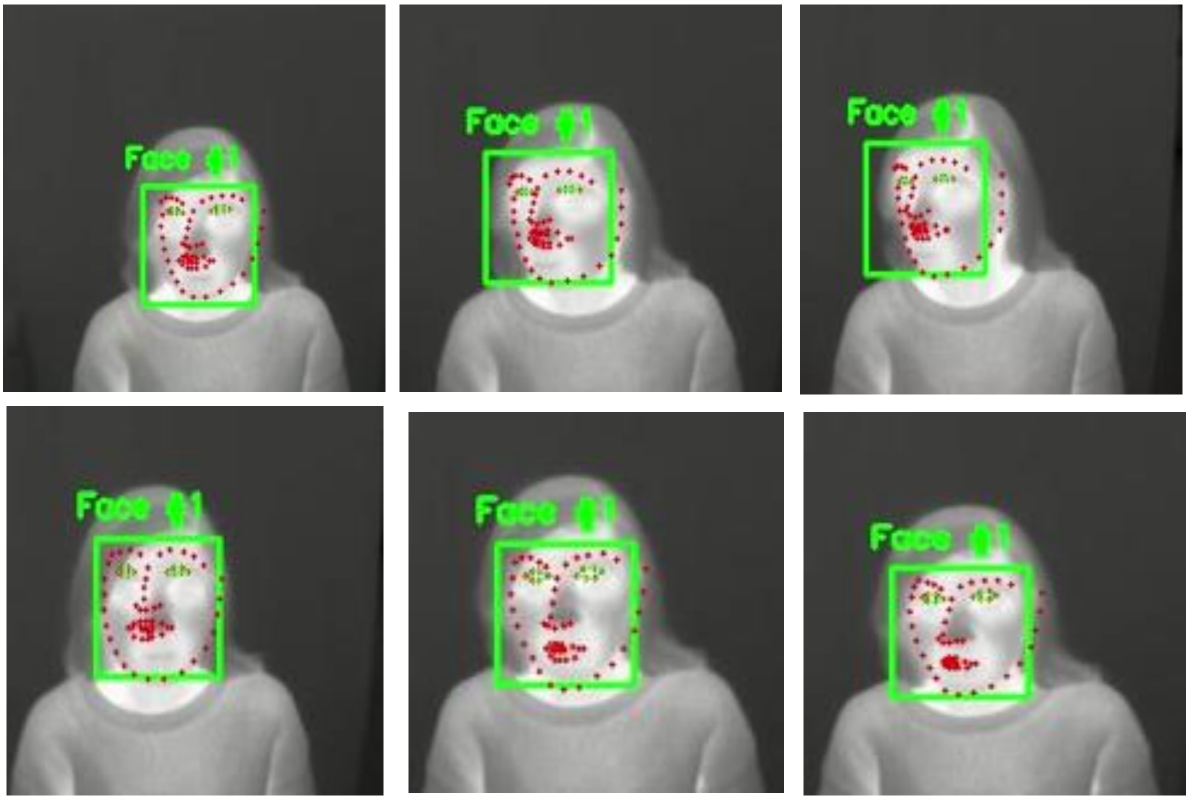


Figure 6: Six different frames showing the output of Dlib based 68 facial landmark detection algorithm. The results are validated on six different facial angles with nearly 5-15 degree angle variation and varying distances of the subject's face from the camera.

4. Conclusions

In this work, we have mainly focused on using conventional computer vision algorithms for face localization and landmark detection on thermal data for in-cabin applications. These types of systems can be further used to extract facial thermal information for drowsiness detection. The performance was tested on locally gathered test data. However, we can further improve the efficacy of these algorithms by training and fine-tuning them on large-scale thermal data as future research work.