



Provided by the author(s) and University of Galway in accordance with publisher policies. Please cite the published version when available.

Title	Detection and characterisation of partially occluded vulnerable road users
Author(s)	Gilroy, Shane
Publication Date	2023-10-29
Publisher	NUI Galway
Item record	http://hdl.handle.net/10379/17930

Downloaded 2024-04-27T20:49:36Z

Some rights reserved. For more information, please see the item record link above.



Detection and Characterisation of Partially Occluded Vulnerable Road Users

A dissertation presented

by

Shane Gilroy

to

The College of Science and Engineering

in fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Electrical and Electronic Engineering

University of Galway

Galway, Ireland

June 2023

Contents

Title Page	i
Table of Contents	ii
Abstract	v
List of Figures	vii
List of Tables	ix
Acknowledgments	x
Glossary of Terms	xi
Declaration of Originality	xii
Sponsor Acknowledgement	xiii
1 Introduction	1
1.1 Motivation	1
1.2 Research Opportunities Targeted	3
1.3 Contributions	5
1.4 Thesis Structure	6
1.5 Publications to date	7
2 Literature Review	9
2.1 Summary	9
2.2 Introduction	10
2.3 Occlusion Handling by Humans	11
2.4 Occlusion Reasoning	15
2.4.1 Region Cues	15
2.4.2 Boundary Cues	16
2.4.3 3D Surface Cues	16
2.4.4 Depth Cues	17
2.5 Occlusion Handling in Pedestrian Detection	17
2.5.1 Segmentation and Parts Based Models	18

2.5.2	Occlusion-Specific Classifiers	21
2.5.3	Pedestrian Detection in Crowds	22
2.5.4	Tracking and Prediction	23
2.5.5	Sensor Fusion and V2X	26
2.5.6	Evaluating Detection Performance of Occluded Pedestrians . .	28
2.5.7	Quantifying Pedestrian Visibility	31
2.6	Occlusion Handling in Vehicle Detection and Tracking Applications .	36
2.6.1	Vehicle to Vehicle Occlusion	36
2.6.2	Two Wheeled Vehicles and Cyclist Detection	41
2.7	Occluded Objects and Traffic Signs	44
2.8	Conclusions and Remaining Challenges	46
3	An Objective Method for Pedestrian Occlusion Level Classification	50
3.1	Summary	50
3.2	Introduction	51
3.3	Related Work	53
3.4	Methodology	53
3.4.1	Occluded Keypoint Detection	56
3.4.2	2D Body Surface Area Estimation	56
3.5	Validation	57
3.5.1	Quantitative Validation	58
3.6	Discussion and Analysis	62
3.6.1	Challenging Image Frames	65
3.7	Conclusions	67
4	The Impact of Partial Occlusion on Pedestrian Detectability	69
4.1	Summary	69
4.2	Introduction	70
4.3	Related Work	72
4.4	Methodology	74
4.4.1	Pedestrian Detection Models	75
4.4.2	Experiments	78
4.5	Results and Analysis	78
4.5.1	Benchmark Comparison	83
4.5.2	Key Semantic Parts	86
4.6	Conclusion	88

5	E-Scooter Rider Detection and Classification in Dense Urban Environments	91
5.1	Summary	91
5.2	Introduction	92
5.3	Related Work	94
5.4	Methodology	97
5.4.1	E-Scooter Rider Classification	100
5.4.2	Occlusion-Aware E-Scooter Rider Detection	100
5.4.3	Performance Characterisation	104
5.5	Results and Analysis	104
5.6	Conclusion	109
6	Conclusions and Future Work	111
6.1	Project Summary and Conclusions	111
6.2	Primary Contributions	116
6.3	Future Work	117
Bibliography		120

Detection and Characterisation of Partially Occluded Vulnerable Road Users

Abstract

Accurate detection and classification of vulnerable road users (pedestrians, cyclists, and micro-mobility users) is a safety critical requirement for the deployment of autonomous vehicles in heterogeneous traffic. Object detection systems have improved significantly in recent years with the proliferation of deep learning-based solutions and the availability of larger and more diverse datasets. Despite this, many challenges still exist before the detection capabilities required for safe autonomous driving can be achieved. One of the most complex and persistent challenges is that of partial occlusion, where a target object is only partially available to the sensor due to obstruction by another foreground object. The frequency and variety of occlusion in the automotive environment is large and diverse as pedestrians, e-scooter riders and cyclists navigate between vehicles, buildings, traffic infrastructure and other road users. Vulnerable road users can be occluded by static or dynamic objects, may inter-occlude (occlude one another) such as in crowds, and self-occlude - where parts of a pedestrian or cyclist overlap. This thesis provides in-depth analysis into this complex object detection challenge and makes significant contributions to the field of research for partially occluded vulnerable road user detection.

The research identifies a number of knowledge gaps and provides advanced characterisation tools to improve the analysis of state of the art pedestrian and e-scooter rider detection models. A thorough literature review of occlusion handling techniques for vehicle detection, vulnerable road user detection and object detection in the automotive environment is presented. A novel, objective metric and methodology for pedestrian occlusion level classification for ground truth annotation is described that more accurately reflects the pixel wise occlusion level than the current state of the art. Two novel, objective test datasets are presented for benchmarking pedestrian

and e-scooter rider detection performance for the complete range of occlusion levels from 0-99%. Finally, a novel occlusion-aware method of e-scooter rider detection is described that provides a 15.93% improvement over the current state of the art.

List of Figures

2.1	Human Occlusion Handling Test Images	12
3.1	Occlusion Level Classification Overview	54
3.2	Occlusion Level Classification Pipeline.	55
3.3	2D Body Surface Area.	58
3.4	Qualitative Validation Results.	59
3.5	Quantitative Validation Dataset Sample Images	61
3.6	Quantitative Evaluation Results 1	63
3.7	Quantitative Evaluation Results 2	64
3.8	Examples of Challenging Image Frames.	66
4.1	Occluded Pedestrian Dataset Statistics	75
4.2	Occluded Pedestrian Dataset Sample	76
4.3	Detection Performance by Occlusion Level	79
4.4	True Positives, False Positives and False Negatives	80
4.5	Overall Detection Performance	81
4.6	FasterRCNN vs. SSD	83
4.7	Example Images from the KITTI Vision Benchmark	85
4.8	Detection Performance by Occlusion Level KITTI Vision Benchmark	86
4.9	Impact of Head Visibility on Detection	87
5.1	E-Scooter Test Dataset Statistics.	98
5.2	E-Scooter Test Dataset Sample.	99
5.3	Candidate Selection Output Comparison.	102
5.4	Occlusion-Aware E-Scooter Detection Flowchart.	103
5.5	Detection and Classification Performance.	105
5.6	Classifier Comparison using the Proposed Occlusion-Aware Pipeline.	106
5.7	Detection Performance by Occlusion Level.	107

5.8	False Negatives by Occlusion Level.	108
5.9	Number of False Positives by Occlusion Level	108

List of Tables

2.1	KITTI High Performing Pedestrian Detection Algorithms	29
2.2	CALTECH High Performing Pedestrian Detection Algorithms	30
2.3	Categories of Occlusion Levels by Dataset.	32
2.4	KITTI High Performing Vehicle Detection Algorithms	41
2.5	KITTI High Performing Cyclist Detection Algorithms	43
3.1	Keypoints to Body Surface Area	60
4.1	Overview of Pedestrian Detection Models.	77

Acknowledgments

I would like to express my sincere gratitude to the following people, without whom this research could not have been completed:

- My primary supervisor Prof. Martin Glavin, co-supervisor Prof. Edward Jones, and contributor Dr Darragh Mullins, for their guidance, expertise, and level-headed advice throughout the research project.
- My Graduate Research Committee at the University of Galway, Prof. Gearóid Ó Laighin, Prof. James Duggan and Prof. Mark Healy for their annual monitoring of the research progress.
- All members of the Connaught Automotive Research Group who provided feedback at various stages of the project.
- My colleagues at IT Sligo and Atlantic Technological University for providing the opportunity and flexibility to conduct my PhD.
- Marty Gilroy, Sean Mullery and Eva Murphy for concurrently suffering through their own PhDs and providing a welcome avenue to vent.
- Josie McKelvey, Tony Gilroy and Bried Gilroy for inadvertently setting me on this path many years ago.
- Most of all, my extremely supportive family Rebecca, Cathal and Aoibhín who kept me on track with weekly check-ups of “Is the PhD done yet?...” and to remind me that there are many more important things in life than a PhD.


Thank you all for the feedback, contributions and support you have provided over the past six years.

Glossary of Terms

ADAS	Advanced Driver Assistance Systems
AHP	Amodal Human Perception
AP	Average Precision
BSA	Body Surface Area
CSRDCF	Channel Spatial Reliability for Discriminative Correlation Filter
DPM	Deformable Part Model
ECO	Efficient Convolution Operators
EKF	Extended Kalman Filter
HOG	Histogram of Orientated Gradients
HVS	Human Visual System
ITS	Intelligent Transportation Systems
LBP	Local Binary Patterns
LSTM	Long Short-Term Memory
mAP	Mean Average Precision
OTA	Object Tracking Accuracy
RNN	Recurrent Neural Network
ROI	Region of Interest
SAE	Society of Automotive Engineers
SVM	Support Vector Machine
TLD	Tracking, Learning and Detection
V2V	Vehicle to Vehicle
V2X	Vehicle to X
VRU	Vulnerable Road User

Statement of Originality

I hereby declare that the work contained in this thesis has not been submitted by me in the pursuance of any other degree.

Name: 
Date: 29 September 2023

Sponsor Acknowledgement

This research has been part funded by Atlantic Technological University.

Chapter 1

Introduction

1.1 Motivation

Approximately 1.3 million people die each year as a result of road traffic incidents according to the World Health Organisation [1]. Over half of all road traffic deaths are vulnerable road users such as pedestrians, e-scooter riders and cyclists [1,2]. Accurate detection and classification of vulnerable road users is a safety critical requirement for the deployment of autonomous vehicles in heterogeneous traffic. The SAE J3016 standard [3,4] defines levels of driving automation ranging from level 0, where the vehicle contains zero automation and the human driver is in complete control, to level 5 where the vehicle is solely responsible for all perception and driving tasks in all scenarios. Level 1 and level 2 automation provide features such as cruise control and lane-keeping assistance to supplement the human driver in limited

scenarios. The progression from automation levels 3-5 represents a significant increase in assumption of responsibility by the vehicle, placing progressively increasing demands on the performance of object detection systems. In level 3 automation, the driver is primarily responsible for the vehicle and automated driving is used in controlled circumstances only, such as highway driving. The human driver is required to take control at any point where more complex driving tasks are encountered. For level 4 automation, the vehicle may encounter more complex driving situations such as suburban or urban scenes, mixing traffic containing other vehicles and vulnerable road users, where the occurrence of significant occlusions may be more likely. Level 5 automation puts the onus on the vehicle to resolve the detail in every scene. The vehicle will be expected to navigate the most densely populated and complex situations, with significant numbers of moving objects (VRUs and vehicles), possibly even in exceptional circumstances where navigation may require interpretation/bending of the rules of the road (e.g. navigating around a road accident, road works or emergency vehicles) to ensure safe passage.

Object detection systems have improved significantly in recent years with the proliferation of deep learning-based solutions and the availability of larger and more diverse datasets. Despite this, many challenges still exist before the detection capabilities required for safe autonomous driving can be achieved. One of the most complex and persistent challenges is that of partial occlusion, where a target object is only partially available to the sensor due to obstruction by another foreground

object. The frequency and variety of occlusion types in the automotive environment is large and diverse as pedestrians, e-scooter riders and cyclists navigate between vehicles, buildings, infrastructure and other road users. Vulnerable road users can be occluded by static or dynamic objects, may inter-occlude (occlude one another) such as in crowds, and self-occlude - where parts of a pedestrian or cyclist overlap. The recent emergence of e-scooter riders further highlights the importance of precise classification of partially occluded road users. E-scooter riders share a large percentage of visual characteristics with pedestrians, however, demonstrate a very different dynamic profile and can reach speeds of up to 45 kilometres per hour. Robust detection and classification is required in order to appropriately inform path planning and accident mitigation in driver assistance and autonomous vehicle applications.

This thesis provides in-depth analysis into this complex object detection challenge and makes significant contributions to the field of research for partially occluded vulnerable road user detection.

1.2 Research Opportunities Targeted

The thesis identifies a number of outstanding knowledge gaps in the field of partially occluded vulnerable road user detection. The specific research opportunities targeted are summarised below.

1. Popular object detection benchmarks such as [5–9] indicate that recent high performing vehicle detection algorithms are commonly able to detect approxi-

mately 90% of partially occluded and 80% of heavily occluded vehicles. However, only 65%-75% of vulnerable road users such as pedestrians and cyclists are detectable under partial and heavy occlusion. A significant amount of research is required to improve the detection of partially occluded pedestrians, cyclists and e-mobility users in the automotive environment.

2. A number of current pedestrian detection benchmarks provide annotation labels for partial occlusion to assess algorithm performance in these scenarios, however each benchmark varies greatly in their definition of the occurrence and severity of occlusion. In addition, current occlusion level annotation methods contain a high degree of subjectivity by the human annotator. This can lead to inaccurate or inconsistent reporting of an algorithm's detection performance for partially occluded pedestrians, depending on which benchmark is used. An objective metric and methodology for pedestrian occlusion level classification is required for ground truth annotation.
3. Current pedestrian detection benchmarks typically categorise occluded pedestrians into two to three broad categories such as "partially" and "heavily" occluded. In addition, many pedestrian instances are impacted by multiple inhibiting factors that contribute to non-detection such as object scale, distance from camera, lighting variations and adverse weather. A detailed, objective benchmark specifically for partially occluded pedestrian detection is required that can be used to objectively characterise detection performance for a com-

prehensive range of occlusion levels.

4. Although similar in physical appearance to pedestrians, e-scooter riders demonstrate distinctly different characteristics of movement and can reach speeds of up to 45kmph. The challenge of detecting e-scooter riders is exacerbated in urban environments where the frequency of partial occlusion is increased as riders navigate between vehicles, infrastructure and other road users. This can lead to the non-detection or mis-classification of e-scooter riders as pedestrians, providing inaccurate information for accident mitigation and path planning in autonomous vehicle applications. Further research on the novel field of e-scooter rider detection is required as it is currently underrepresented in vulnerable road user detection benchmarks.

1.3 Contributions

The primary contributions of this thesis can be summarised as follows:

- A comprehensive literature review on the theme of occluded object detection in the automotive environment as published in IEEE Transactions on Intelligent Transportation Systems (2019) [10].
- A novel, objective metric and methodology for pedestrian occlusion level classification for ground truth annotation as published in ICCV Workshop on

Occluded Video Instance Segmentation (2021) [11] and Pattern Recognition Letters (2022) [12].

- A novel, objective, test benchmark for partially occluded pedestrian detection as published in Biomimetic Intelligence and Robotics (2023) [13].
- A novel, objective, test benchmark for partially occluded e-scooter rider detection and classification as published in Results in Engineering (2022) [14].
- A novel, occlusion-aware method of e-scooter rider detection is proposed that provides a 15.93% improvement over the current state of the art, as published in Results in Engineering (2022) [14].

1.4 Thesis Structure

The remainder of this thesis is structured as follows: **Chapter 2** provides a thorough literature review of the current state of the art methods for detecting partially occluded pedestrians, vehicles (including two wheeled vehicles and cyclists) and objects in the automotive environment. **Chapter 3** describes an objective metric and methodology for quantifying and annotating the severity of occlusion for partially occluded pedestrians. **Chapter 4** analyses the impact of partial occlusion on pedestrian detectability and characterises the performance of popular pedestrian detection models across a range of occlusion levels from 0-99% occluded. **Chapter 5** investigates the impact of occlusion on e-scooter rider detection and proposes a

novel method of e-scooter rider detection that provides a 15.93% improvement in detection performance compared to the current state of the art. **Chapter 6** outlines the conclusions, recommendations and further research opportunities identified throughout this research project.

1.5 Publications to date

Four journal papers and one conference paper have been submitted for publication from this work:

Journal Papers

- *S. Gilroy, E. Jones, and M. Glavin, “Overcoming occlusion in the automotive environment-a review”, IEEE Transactions on Intelligent Transportation Systems, 2019. [10]*
- *S. Gilroy, M. Glavin, E. Jones, and D. Mullins, “An objective method for pedestrian occlusion level classification”, Pattern Recognition Letters, 2022. [12]*
- *S. Gilroy, D. Mullins, A. Parsi, E. Jones, and M. Glavin, “Replacing the human driver: An objective benchmark for occluded pedestrian detection”, Biomimetic Intelligence and Robotics, 2023. [13]*

- *S. Gilroy, D. Mullins, E. Jones, A. Parsi, and M. Glavin, “E-scooter rider detection and classification in dense urban environments”, Results in Engineering, Vol.16, 2022. [14]*

Conference Papers

- *S. Gilroy, M. Glavin, E. Jones, and D. Mullins, “Pedestrian occlusion level classification using keypoint detection and 2d body surface area estimation”, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3833–3839. [11]*

Chapter 2

Literature Review

2.1 Summary

Accurate and consistent vulnerable road user detection remains one of the most challenging perception tasks for autonomous vehicles. One of the most complex outstanding issues is partial occlusion, where a sensor has only a partial view of the target object due to a foreground object that partially obscures the target. A review of occlusion detection and handling solutions for the automotive environment is presented in this chapter. The literature review first discusses object detection by the human visual system, provides an overview of occlusion reasoning in computer vision, presents a summary of occlusion handling strategies in pedestrian, vehicle and object detection applications in the automotive environment. A selection of the remaining challenges to achieving the required level of object detection performance

for safe autonomous driving are also discussed.

2.2 Introduction

Vision systems have become integral for road user detection in driver assistance applications however many challenges still exist before the object detection capabilities required for safe autonomous driving are reached. One of the most challenging outstanding issues is occlusion, where a target object is only partially available to the sensor due to obstruction by another foreground object. Occlusion exists in various forms ranging from partial occlusion to heavy occlusion. In the automotive environment, target objects can be occluded by static objects such as buildings and lampposts, dynamic objects such as moving vehicles or other road users, may inter-occlude (occlude one another) such as in crowds, and self-occlude where parts of a pedestrian or cyclist overlap. The frequency and variation of occlusion in the automotive environment is vast and can also be impacted by cultural and environmental factors. Current benchmarks such as [5–7] indicate that recent high performing vehicle detection algorithms are commonly able to detect approximately 90% of partially occluded and 80% of heavily occluded vehicles, however only 65%-75% of vulnerable road users such as pedestrians and cyclists are detectable under partial and heavy occlusion. This research provides an overview of state-of-the-art occlusion detection and handling methods in the automotive environment.

The remainder of this chapter is organised as follows: Section 2.3 provides an

overview of analysis on the human approach to identifying and recognising partially occluded objects. Section 2.4 provides an overview of occlusion reasoning and outlines multiple cues that can be used to identify cases of occlusion in computer vision. Section 2.5 provides an overview of occlusion handling strategies for pedestrian detection applications. Section 2.6 provides an overview of occlusion handling strategies for vehicle detection and tracking applications, including two-wheeled vehicles and cyclists. Section 2.7 provides an overview of occlusion handling strategies for objects and traffic signs. Section 2.8 discusses a selection of the remaining challenges to improving the detection of partially occluded objects.

2.3 Occlusion Handling by Humans

The human vision system (HVS) has an adept ability to recognise objects under partial occlusion. Research suggests that humans can differentiate between complex visual categories within 150ms of been provided with the stimulus [15–17]. Stereopsis, the depth perception capabilities granted by the binocular nature of human vision [18], and the process of Amodal Completion provide humans with distinct advantages when identifying partially occluded objects. Amodal Completion allows humans to perceive objects as a whole, despite partially occluded or missing information, through the continuation or inference of contours in a scene [19–22]. The identification of object parts and their known spatial relationship also inform the presence of an object under partial occlusion. The salience of individual parts is

determined by three factors: the protrusion, boundary strength and relative size of the part [23]. A wide range of psychophysical studies have been carried out in an attempt to understand the methods by which humans detect and recognise occluded objects in complex scenes.

Fukushima [24] outlined an experiment to investigate human ability to recognise objects through partial occlusion. The author challenged participants to recognise letters of the alphabet under two forms of distortion, Figure 2.1.

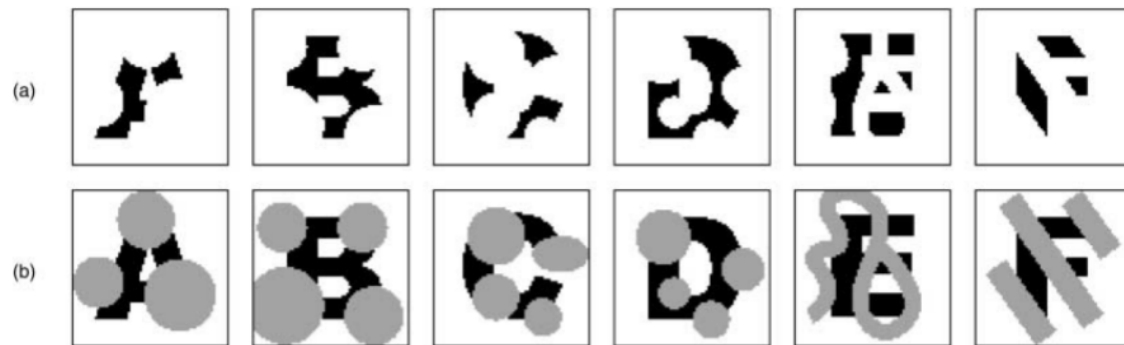


Figure 2.1: Test images used by Fukushima to investigate the human approach to occlusion handling. (a) Sections of each letter omitted; (b) Sections of each letter under visible occlusion.

This study observed that the participants required more time to recognise a known pattern that was incomplete or had sections missing, than to detect the same pattern under visible occlusion. A neural network model was constructed based on findings from this study to emulate the human approach to this task. It was concluded that at the recognition stage, the model can easily identify the object features if the occluding object is visible, however it struggled to distinguish which features belong to the original pattern if the occluded section is removed. The

findings of this study are corroborated by Johnson and Olshausen [25] who carried out a series of three experiments using an electroencephalography (EEG) to analyse human recognition time of partially visible objects in the presence and absence of their occluding patterns. Again, the author concludes that participants were able to identify partially occluded objects faster when the occluding pattern is visible. Meng and Potter [26] conducted a range of experiments to investigate how occlusion impacts human object detection and recognition. The study observed that humans are better at detecting or searching for a known target under partial occlusion, than recognising objects that had been partially occluded in an image after the event. The results of these experiments also indicate that contextual “gist” information improves the human visual systems ability to tolerate noise.

Struwe [27] investigated human object recognition of occluded objects through the use of an eye tracking device. The results of this study suggest that participants first identify the boundary and profile of the occluding object to inform a more localised search for the occluded object. The author concludes that humans use the presence and details of the occlusion to aid detection of occluded objects and that a hierarchical approach of isolating easy to detect objects first, then looking for the more difficult ones may increase the performance of object detection in computer vision applications. A recent neurophysiological study by Fyall et al [28] proposes that the visual cortex, the section of the brain which carries out initial processing of visual information from the retina, is proficient at identifying simple objects. How-

ever, in more complex tasks such as identification of objects under partial occlusion, communication occurs between the visual cortex and the prefrontal cortex, a higher brain region involved in memory and learning. Several studies have been carried out to develop an accurate model of how the human brain recognises visual objects. Many proposals suggest that this may be explained through a feedforward process [29–31] and can be represented in computer vision through the implementation of a Feedforward Deep Neural Network [32, 33]. This feedforward approach to object recognition becomes challenged when tasked with identifying objects under occlusion however, as occluding an object can underspecify a stimulus, so that it initially does not resemble stored patterns. The use of feedback or recurrent loops in these cases can ultimately recover the correct pattern by reinforcing the initial weak image representation for the duration of the stimulus. Wyatte et al [34], O’Reilly et al [35] and Spoerer et al [36] concluded that Recurrent Convolution Neural Networks are a more suitable model for biological recognition of occluded objects.

Despite significant improvements in recent years, synthetic and engineered models still fall short of human performance for object detection in the automotive environment. Zhang et al [37, 38] determined from experiments carried out using the Caltech [6] and KITTI [7] pedestrian detection benchmarks, that there is a tenfold gap in terms of errors to be closed between current technology and a human baseline. The largest deficiency of current detection models is their difficulty in identifying small scale and partially occluded targets.

2.4 Occlusion Reasoning

In order to robustly handle occlusions in the automotive environment, object detection systems may first successfully identify cases of occlusion within the field of view. One popular method of such occlusion reasoning is through the analysis of motion cues such as frame comparison reasoning. Frame comparison reasoning determines occlusion by analysing continuous image data and identifies objects by comparing data between frames. Motion Cues are very effective in detecting and tracking objects where the level of occlusion varies with time, however, it is restricted in cases of static occlusion where variation of occlusion between frames is reduced. Many other popular methods of occlusion reasoning combine a number of the following occlusion cues or image characteristics to assess if an object boundary is due to an occlusion and if so, if it is more likely to be the occluding or occluded object [39–46].

2.4.1 Region Cues

Adjacent artefacts are likely to be different objects with an occlusion boundary between them if they have different colors, textures or are misaligned. Image position can also be used to inform occlusion reasoning as lower regions in an image tend to be closer to the camera or sensor [43].

2.4.2 Boundary Cues

Long, smooth boundaries with strong color or texture gradients are more likely to be occlusion boundaries than short boundaries with weak gradients. The geometry of the boundary can also provide useful information about the presence and nature of an occlusion; this is particularly the case with the convexity of a shared boundary. If a shared boundary appears to be convex it is likely that one region occludes another [43].

2.4.3 3D Surface Cues

3D surface characteristics can be used to identify many occlusion boundaries or differences in adjacent regions. Hoiem et al [39] illustrated this by using the example of a pedestrian in front of a building. Analysing the 3D surface characteristics in this example identifies a non-planar surface (the pedestrian) occluding a planar horizontal surface (the road/footpath) and a planar vertical surface (the wall of the building). Occlusion can also be inferred by the presence of junctions, particularly T-Junctions at the boundaries of surfaces [39, 45, 47]. T-Junctions are formed by three regions, one of which forms an almost flat angle of approximately 180° and two other regions of arbitrary angles, but generally angles of more than $30\text{-}40^\circ$ are required to maintain the perception of occlusion [48].

2.4.4 Depth Cues

A boundary is likely to be an occlusion boundary if there is a large depth discontinuity between adjacent regions. Absolute depth information can be easily obtained in lidar and stereo vision applications. Relative depth between objects can be estimated in monocular vision systems by determining where each region contacts the ground. In cases where the ground contact for a region is occluded, possible relative depth can be estimated based on the visible portion of the region and its occlusion relationships to surrounding objects with known depth [39]. Convexity cues can also be used to infer the relative depth information of occluding objects i.e. convex shapes appear to be in front of their background etc. [43, 45, 46]. Rezaeirowshan et al [46] propose a bio-inspired method of extracting global depth order from a single image using monocular depth cues. This method determines local depth order between adjacent shapes by analysing the convexity of shared boundaries, then detects and analyses T-Junctions in the image to obtain depth order between shapes. Approximation of rank aggregation is then used to establish global depth order from the local cues.

2.5 Occlusion Handling in Pedestrian Detection

Detection, classification and tracking of pedestrians in real world scenarios can be particularly challenging due to their tendency to deform, self-occlude and inter-

occlude. The introduction of AlexNet [32] in 2012 led to a significant improvement in pedestrian detection systems through the use of Convolutional Neural Networks (CNN). AlexNet achieved a top-5 error of 15.3%, more than 10.8% lower than the next best network at the ImageNet Large Scale Visual Recognition Challenge 2012 [49]. The breakthrough triggered a renewed focus on the development and improvement of deep learning-based solutions for pedestrian detection with progressively increasing results. A thorough roadmap on the evolution of object detection models over the past 20 years can be found in [50, 51]. Despite recent achievements in deep learning based pedestrian detection routines, partial occlusion has been consistently highlighted as one of the most complex outstanding pedestrian detection challenges in research survey papers from 2009 [5] through to 2023 [10, 52, 53]. This section provides an overview of the occlusion handling strategies used in pedestrian detection applications.

2.5.1 Segmentation and Parts Based Models

A popular approach to occlusion handling is to divide a target image or region of interest into a number of cells or segments and then analyse each segment individually to improve detection results. Parts based models divide a target object into semantic or distinguishable parts and classify each part individually to indicate the presence of the target object. Wang et al [54] proposed an occlusion handling technique that merges global and parts-based detection strategies. Classification scores per

cell in a sliding window detection system are calculated by combining a Histogram of Oriented Gradients (HOG) based detector with Local Binary Patterns (LBP) in order to improve detection performance under partial occlusion. Occlusion likelihood maps are constructed using the response of each block of the HOG feature to the global detector. The sum of the HOG block responses to the global detector indicates possible partial occlusion. Parts-based detectors (upper body & lower body) are then applied to the unoccluded regions to perform classification of occluded pedestrians. This approach was evaluated further by Het Veld et al [55] which determined that the most significant aspect of the technique is the integration of partial classifiers, rather than the negligible impact of the cell based occlusion detection and region merging. Gao et al [56] proposed a similar strategy in which a set of binary variables are attached to each cell indicating if the pixels in each cell belong to the object. This method uses a structural Support Vector Machine (SVM) to learn the values of the binary variables. However, the technique requires the model to be trained with features of both the occluding and occluded object. Chen et al [57] also used a similar strategy to reduce the negative impact of sunglasses or scarves on facial recognition. Occluded regions are identified by dividing the facial image into six symmetrical patches and examining each patch separately. Patches containing occlusions are then removed, and classification is carried out on the remaining area.

Parts-based methods for pedestrian detection are often more robust than holistic models in terms of occlusion handling, as they accumulate the detection responses of

the visible features of partially occluded pedestrians. The loss of score of a possible object hypothesis is proportional to the severity of the occlusion [27, 55, 58]. Chan et al [59] proposed a method for detecting partially occluded pedestrians by determining the visible parts of the object. This research uses a discriminatively trained Deformable Part Model (DPM) containing binary visibility flags which are used to indicate whether an image section belongs to the target object or the occluder by solving a concave optimisation problem. Occluding sections are then removed from the classification process so that only the visible regions of the target object are calculated. Optimisation is carried out to identify occluded sections using the following known characteristics of pedestrian occlusion in traffic scenes: (i) partial occlusions occur more often in the lower part of the body than the upper part of the body [6] and (ii) occluded pixels tend to be spatially clustered and form connected regions. Ouyang et al [60–62] similarly used a deformable part based model and hidden occlusion variables in order to identify and remove the impact of the occluder from classification, however in this case a discriminative deep model is used to learn the visibility relationships of overlapping parts at multiple layers. Baumgartner et al [63] used stereo images and depth maps to identify Regions of Interest (ROIs) by projecting 3D point clouds onto a ground plane. The ROIs were then segmented into individual object areas separating the occluded from the occluding objects. This information was then used for tracking and classification.

Enzweiler et al [44] identified the occluding object by incorporating information

gained from motion, depth and segmentation results. Each test sample was segmented with depth and motion cues to determine occlusion component weights. The detection confidence scores for different parts were used to estimate their visibilities and were computed as weighted means of multiple cues for different parts. The primary occlusion handling strategy of segmentation or parts-based models is often to identify and remove an occluding object in order to increase the classification score of the occluded object. This process reflects the human visual system’s approach to classifying partially occluded objects as identified in [24, 25, 27].

2.5.2 Occlusion-Specific Classifiers

Another widely used approach for improving occlusion detection and handling is to train occlusion-specific classifiers. Kwak et al [64] trained a single occlusion-specific classifier based on observation likelihoods. The target was divided into a regular grid and the state of occlusion of each cell was determined by the classifier. Multiple occlusion-specific classifiers can be merged in order to increase performance [65], particularly for the most frequent types of pedestrian occlusion, occlusions from the bottom, the right or the left [6]. Wojek et al [66] used these trends to train a small number of occlusion-specific classifiers, each for a different occlusion type. Objects were identified using depth information from a monocular camera and the appropriate classifier was applied based on the characteristics of the occlusion, i.e. from the bottom, the left or the right. These approaches can be inefficient as computational

cost and test time increases in a linear fashion with the size of the classifier set [67] and the possible variation of types of occlusion in the real world is vast. Mathias et al [68] proposed a method known as “Franken Classifiers” to increase the efficiency of such techniques by reusing computations among different training stages to reduce the time-cost of training multiple occlusion-specific classifiers. The benefits of both of these techniques converged in Het Veld et al [55] to produce an object detection system with 17 different classifiers, based on varying levels of occlusion, merged with a parts based detection model derived from [54] in order to obtain real time detection. The author claims that this solution can provide an 8% performance improvement in detecting occluded pedestrians versus a single classifier baseline system, while incurring only a 3.4% increase in computational cost.

2.5.3 Pedestrian Detection in Crowds

The challenges posed by occlusion in pedestrian detection can be exacerbated when attempting to identify multiple pedestrians in a crowd. Some approaches train models specifically to handle pedestrian-pedestrian occlusions such as Tang et al [69] who expanded the deformable part model to develop a double person detector, trained with 1,300 synthetically generated images containing various degrees of pedestrian-pedestrian occlusion. The author claims that this approach can outperform single person detectors by more than 25% when faced with images of a pedestrian occluded by one other pedestrian, however, performance is greatly reduced

when identifying single pedestrians. Fusion of the double person detection model with a single person detector can greatly improve the performance of pedestrian detection in images of multiple occluding people. A similar approach is proposed in [70] which fused a single pedestrian detector with a specifically trained multi-pedestrian detector and used a tailored probabilistic framework to model the configuration relationship between the single and multi-pedestrian detectors.

2.5.4 Tracking and Prediction

Tracking, prediction and frame comparison models can be used to handle occlusions by monitoring and matching activity between frames, i.e. if an identified object is tracked over multiple frames and then becomes occluded or disappears for a period, prediction models can be used to estimate the proposed location of the object until it reappears. Early work in this area [71] used an Extended Kalman Filter (EKF) to estimate trajectory based on velocity and position. When occlusion occurs, the filter provides the maximum likelihood estimate of the occluded region. Occlusion can be predicted in tracking routines by monitoring the frames leading up to the occlusion or merging of blobs/binary maps and thresholds can be set for implementing occlusion management routines or flagging that an occlusion may have occurred. The merged blob can then be monitored while an occlusion is flagged to limit the area of the occlusion. It can then be expected that the tracked/identified blob will eventually separate providing strong evidence for the end of an occlusion.

The approximate duration of the occlusion can be predicted using the EKF estimates of the 3D velocity and position along with the calculated area of the blob. This is also referred to as the “merge-split” approach to occlusion handling [72]. A number of challenges can occur when attempting to associate a tracked object following occlusion in these cases. Once a split occurs and an object reappears it can be difficult to assess if this is the same tracked object or a new object introduced to the scene. This can be a particular issue when tracking pedestrians or other objects that have a similar appearance. A number of proposed solutions, known as occlusion recovery methods are presented in [42, 73, 74].

Vethamani and Diala [75] presented a spatio-temporal tracking approach to handling both partial and full occlusions in tracking applications. The severity of occlusion was determined using histogram matching and edge detection to calculate the area of the tracked artefact. A threshold was then set for the area to be calculated. The size of this threshold indicates if an artefact is partially occluded and a spatial approach should be used, or if it is fully occluded and a temporal occlusion handling routine is required. In the case of full occlusion, the reference frame is compared with previous frames based on texture to identify and fill in the missing pedestrian outline and display the result. In the case of partial occlusion, edge-based restoration is used. Edges are detected, and features are extracted around the artefact based on contours. The image can then be reconstructed by inventing content based on known image properties. Sadeghian et al [76] tracked multiple targets using a structure of

Recurrent Neural Networks (RNN) that learned to encode long term dependencies from a combination of appearance, motion and interaction cues. Motion and interaction models were used to inform Long Short-Term Memory (LSTM) networks and an appearance-based model learned similarity metrics to track targets through long term occlusion.

In 2014 Smeulders et al [77] carried out an assessment of nineteen different tracking strategies and ranked their performance in terms of a calculated Object Tracking Accuracy (OTA). This study found that a Tracking, Learning and Detection (TLD) discriminative classification strategy was the most effective approach for handling occlusion in the test sequences for both static and moving camera applications. TLD, outlined in [77, 78], merges a discriminative classifier and an optical flow tracker. The detector learns an appearance model from the 2-bit binary patterns of the initial bounding box using a Random Fern approach [79]. The algorithm then selects locations with the highest detector scores in each new frame. The optical flow tracker applies a Lucas-Kanade Tracker to map locations to the previous frame in order to propose a target window in each new frame. Normalised cross correlation is then calculated to select the candidate window with the highest similarity to the object model as the new object. This research has been superseded by an updated comparative study of tracking approaches carried out in 2018 by Fiaz et al [80]. Fiaz et al concludes that the Efficient Convolution Operators (ECO) tracking scheme [81] and the Channel Spatial Reliability for Discriminative Correlation Filter (CSRDCF)

tracking algorithm [82] currently display the highest performance for occlusion handling in tracking applications. It was found that the ECO approach was the highest performing overall, however, CSRDCF is slightly more robust in noisy environments [80].

2.5.5 Sensor Fusion and V2X

Kwon et al [83, 84] proposed a Lidar/Radar sensor fusion technique for detecting partially occluded pedestrians in an attempt to reduce the heavy computational requirements and light sensitivity constraints of camera-based systems. Multiple regions of interest (ROIs) are identified by lidar and radar sensor measurements respectively. Fusion regions of interest are then identified by superimposing the lidar ROIs and radar ROIs. Occlusion ROIs or potential occluded targets are identified by overlapping the occluded depth information obtained from the lidar measurement with the radar ROI, thereby using the positional and doppler information to determine if a moving object exists within the occlusion ROI. While walking, humans produce unique, repetitive Doppler and micro-Doppler patterns as one leg remains fixed as the other takes a step forward [85]. This radar Doppler pattern, which distinguishes humans from other obstacles, is then used to determine if the occluded object is a pedestrian.

The Camera/Lidar fusion based “F-PointNet” algorithm presented by Qi et al [86] uses Frustum PointNets for 3D object detection using RGB-D data. 2D regions

are detected and classified using a 2D CNN, before being extruded into 3D frustum proposals using lidar depth data. A 3D bounding box is then generated from the points in frustum. F-PointNet displays very strong occlusion handling abilities, having achieved an Average Precision (AP) of 77.25% on partially occluded and 74.46% on heavily occluded test data in the KITTI Dataset for pedestrian detection. Many similar strategies such as [87–91] proposed Camera/Lidar fusion to improve the detection of partially occluded pedestrians.

Huang and Jiang [92] fused a color camera with a thermal imaging camera to improve pedestrian detection and tracking through occlusion. This approach can track pedestrians in scenarios where the target is severely occluded in the color image however is still emitting thermal radiation. Variations in thermal emission due to clothing can also be used to inform segmentation and tracking in cases of pedestrian to pedestrian occlusion [93]. Bo et al [94] presented a multi-camera fusion method for pedestrian detection and tracking which takes into account the level of occlusion computed from the projected geometry in each viewpoint and dynamically attaches a weighting to the viewpoint with the lowest level of target occlusion. Vehicle to Vehicle (V2V) or Vehicle to X (V2X) communications [95, 96] allow vehicles and intelligent infrastructure (traffic light, traffic sign etc.) to communicate the presence of detected vulnerable road users to other vehicles which may not have a clear line of sight, therefore potentially mitigating the severity of occlusions in well-instrumented urban environments.

2.5.6 Evaluating Detection Performance of Occluded Pedestrians

The KITTI Vision Benchmark Suite [7] provides an opportunity to assess how detection models compare in terms of occlusion handling. KITTI consists of 7,481 training images and 7,518 test images containing 80,256 labelled objects, including pedestrian, car, bicycle and occlusion-specific annotations. Images in the KITTI Dataset are divided into three levels of difficulty for assessment purposes:

- *Easy* - Minimum Bounding Box Height of 40 Pixels, Maximum Occlusion Level: Fully Visible, Maximum Truncation: 15%
- *Moderate* - Minimum Bounding Box Height of 25 Pixels, Maximum Occlusion Level: Partly Occluded, Maximum Truncation: 30%
- *Hard* - Minimum Bounding Box Height of 25 Pixels, Maximum Occlusion Level: Difficult to see, Maximum Truncation: 50%

Assessing algorithm performance on the “Moderate” and “Hard” test data can provide an indication of the comparative performance of detection models for partially occluded and heavily occluded objects respectively. Results on the KITTI benchmark suite are displayed in Average Precision (AP), Table 2.1. Average Precision is a popular object detection metric described in [97]. It is calculated by plotting the precision-recall curve for a detection model and then detecting the area under the curve (AUC).

Table 2.1: KITTI High Performing Pedestrian Detection Algorithms

Performance (Average Precision)			
Algorithm	Moderate (Partial Occlusion)	Hard (Heavy Occlusion)	Detection Model / Strategy
F-PointNet [86]	77.25%	74.46%	<i>Camera Lidar Fusion RGB-D Data, CNN, 2D to 3D Image Extrusion</i>
TuSimple [98, 99]	77.04%	72.40%	<i>CNN Scale-Dependent Pooling, Layer-Wise Cascaded Rejection Classifiers</i>
RRC [7, 100]	75.33%	70.39%	<i>CNN Recurrent Rolling Convolution (RRC)</i>
MS-CNN [15]	73.62%	68.28%	<i>CNN Adaptive CNN based on object scale</i>
GN [101]	71.55%	64.82%	<i>CNN Guiding Network</i>
SubCNN [20]	71.34%	66.36%	<i>CNN Sub-category aware region proposal network</i>

Another commonly used dataset for assessing the performance of detection algorithms for partially and heavily occluded pedestrians is the Caltech Pedestrian Dataset [5, 6]. The Caltech dataset computes results for partially and heavily occluded pedestrians in terms of Log Average Miss Rate (LAMR). LAMR is calculated by first calculating the Miss Rate for the model at a range of confidence thresholds. The Miss Rate is defined as the ratio of false positive detections to the total number of ground truth objects. The log of the Miss Rate is calculated for each class and the LAMR is defined as the average of the logged Miss Rates. In contrast to the KITTI

dataset, the lower the algorithm score on the Caltech dataset indicates the more optimum performance, Table 2.2. Other popular pedestrian detection benchmarks with occlusion specific annotation are KAIST [102], Multi Object Tracking (MOT) [103], Pascal VOC [104], CityPersons [8], Multispectral Pedestrian Dataset [105] and Daimler Multi-Cue Occluded Pedestrian Classification Benchmark [44].

Table 2.2: CALTECH High Performing Pedestrian Detection Algorithms

Performance (Log Average Miss Rate)			
Algorithm	Moderate (Partial Occlusion)	Hard (Heavy Occlusion)	Detection Model / Strategy
SDS-RCNN [106]	15%	59%	CNN Simultaneous Detection and Segmentation using Semantic Feature Information
F-DNN+SS [107]	15%	54%	CNN Multiple Parallel Deep Neural Networks, Pixel-wise Segmentation
F-DNN [107]	15%	55%	CNN Multiple Parallel Neural Networks, Soft-rejection based Fusion
PCN [108]	16%	56%	CNN CNN + Part and Context Information
MS-CNN [15]	19%	60%	CNN Adaptive CNN based on object scale
DeepParts [109]	20%	60%	CNN Parts-based Model

Current popular pedestrian detection benchmarks do not differentiate between

e-scooter riders and pedestrians. This can cause significant issues in autonomous vehicle applications as the detection output is used to inform path planning and accident mitigation. Although partially occluded e-scooter riders can appear very similar to pedestrians from a perception point of view, their dynamic profile and characteristics of movement differ largely as e-scooters can reach speeds up to 45kmph [110–112].

2.5.7 Quantifying Pedestrian Visibility

This section provides an overview of current occlusion level classification methods for pedestrian detection, pedestrian analysis for flood level assessment and commonly used methods for estimating the visibility of pedestrians.

A number of publicly available datasets provide annotation of the level of pedestrian occlusion in the automotive environment. Table 2.3 provides an overview of the categories used to define the severity of occlusion in current popular datasets. Analysis of current benchmarks demonstrate the range of inconsistency and subjectivity in the definition of low, partial and heavy occlusion. The Eurocity Persons Dataset [9] categorises occlusion into three distinct levels: low occlusion (10%-40%), moderate occlusion (40%-80%), and strong occlusion (larger than 80%). Classification is carried out by human annotators. The full extent of the occluded pedestrian is estimated, and the approximate level of occlusion is then estimated to be within one of the three defined categories. This process is also used to classify the level of

Table 2.3: Categories of Occlusion Levels by Dataset.

Dataset	Occlusion Level		
	<i>Low</i>	<i>Partial</i>	<i>Heavy</i>
EuroCity Persons [9]	<40%	40-80%	>80%
CityPersons [8]	-	<35%	35-75%
KITTI [7]	“Fully Visible”	“Partially Occluded”	“Difficult to See”
Caltech Pedestrian [5]	-	1-35%	35-80%
Multispectral Pedestrian [105], OVIS [53]	-	$\leq 50\%$	$>50\%$
TJU-DHD [113]	-	$\leq 35\%$	$>35\%$
Daimler Tsinghua [114]	<10%	10-40%	41-80%
Li <i>et al</i> 2017 [115]	“Fully Visible”	1-40%	41-80%
SAIL-VOS [116]	-	1-25%	>25-75%

truncation of pedestrians near the image border. A similar approach is undertaken in the Caltech Pedestrian [5][38], TJU-DHD-pedestrian [113], CrowdHuman [117] and PedHunter [118] datasets in which pedestrians are annotated with two bounding boxes that denote the visible and full pedestrian extent. In the case of occluded pedestrians, the location of hidden parts of the full pedestrian were estimated by the human annotator in order to calculate the occlusion ratio. Further analysis of the Caltech Pedestrian [5] dataset determined that the probability of occlusion in the automotive environment is not uniform, but rather has a strong bias for the

lower portion of the pedestrian to be occluded and for the top portion to be visible. Classification of occluded pedestrians in the CityPersons dataset [8] is achieved by drawing a line from the top of the head to the middle of the two feet of the occluded pedestrian. Human annotators are required to estimate the location of the head and feet if these are not visible. A bounding box (“ $BB - full$ ”) is then generated for the full pedestrian area using a fixed aspect ratio of 0.41 (width/height). A visible pedestrian area bounding box (“ $BB - vis$ ”) is also annotated and the occlusion ratio is calculated as $Area(BB - vis)/Area(BB - full)$. These estimates of occlusion level are then categorised into two levels in the Citypersons benchmark, Reasonable ($\leq 35\%$ occluded) and Heavy Occlusion (35%-75%). Although this approach can yield plausible results for pedestrians who are standing upright with their arms by their side, the use of a fixed aspect ratio can restrict efficacy in instances with other poses such as crouching, bending over or standing with their arms outstretched.

A more semantic approach to determining the occlusion level was taken in the KITTI Vision Benchmark [7], where human annotators were simply asked to mark each bounding box as “visible”, “semi-occluded”, “fully-occluded” or “truncated”. A similar approach was used in the Multispectral Pedestrian Dataset [105] where pedestrians occluded to some extent up to one half are tagged as partial occlusion; and those whose contour is perceived to be mostly occluded were tagged as heavy occlusion during ground truth annotation. Occluded Video Instance Segmentation (OVIS) [53] estimates the degree of occlusion by calculating the ratio of intersecting areas of

overlapping bounding boxes to the total area of the respective bounding boxes. The authors acknowledge that although this proposed “Bounding Box Occlusion Rate” can be a rough indicator for the degree of occlusion, it can only reflect the occlusion between objects in a partial way and it does not accurately represent the pixel-wise occlusion level of the target objects.

Chaudhary *et al* [119], propose a method of flood level classification from social media images based on the visibility of pedestrians in the image. Assuming the average height of a human adult is estimated to be 170cm, the flood level classifier detects pedestrians in an image and estimates how much of the pedestrian is covered by flood water by vertically subdividing the pedestrian into 11 distinct levels. The highest level of the pedestrian occluded by the water indicates the flood height in the image location. Feng *et al* [120] estimates flood level based on the relative height of specific human body parts which are perceived to be below the water line. The water line in the image is hypothesized to be at the bottom line of the bounding box of a person. A similar approach is taken by Quan *et al* [121] in which keypoint detection is correlated with a binary mask output of a pedestrian detector. Analysis is then carried out to determine if keypoints which represent the hip or knees are outside of the detected binary mask area due to occlusion by flood water in the image, thereby indicating a relative flood level. Noh *et al* [122] approximate the severity of pedestrian occlusion by dividing a pedestrian bounding box into a 6x3 section grid. Detection confidence values are calculated by applying a pedestrian classifier

to grid section and a part confidence map is produced for the complete bounding box. Zhang *et al* [123] assess pedestrian occlusion level by segmenting pedestrians into 5 distinct sections. Each segment is assigned a fixed height and width relative to the total bounding box based on the empirical ratios identified in [124]. ROI pooling is used to detect features within each section and visibility scores are calculated to indicate the relative pedestrian occlusion level.

Wallace [125] proposed a method of classification of body surface area for the purposes of diagnosing the severity of burn damage of the average adult burn victim [126]. This method, known as the “Wallace Rule of Nines”, is commonly used by emergency medical providers and first responders to assess the total affected body surface area of burn patients [127, 128]. The Rule of Nines estimates total body surface area by assigning percentages, in multiples of 9% to semantic body areas, based on the relative physical dimensions of the average adult. The head is estimated to be 9% of the total body surface area (4.5% for the front and 4.5% for the rear). The chest, abdomen, upper back and lower back are each assigned 9%. Each leg is assigned 18%, each arm is assigned a total of 9% and the groin is assigned the remaining 1%. Further research such as [127, 129] validate the Rule of Nines for use in the assessment of total body surface area for the average adult. However, these studies also provide amendments to more accurately reflect body proportions in specific edge cases such as obese adults and infant children.

2.6 Occlusion Handling in Vehicle Detection and Tracking Applications

Accurate detection, tracking and path prediction of partially occluded or emerging vehicles is potentially a significant contributor to mitigating the severity of road accidents and reducing the number of fatal road traffic accidents worldwide. This section provides an overview of occlusion handling strategies used in vehicle detection for advanced driver assistance systems (ADAS), intelligent transportation systems (ITS) and traffic monitoring applications.

2.6.1 Vehicle to Vehicle Occlusion

Yin and Sun [130] proposed an adaptive multi-strategy method for tracking multiple occluding or occluded vehicles. This algorithm first determines the occlusion relationship and the degree of occlusion of each vehicle. Vehicles are then categorised into severe occlusion, partial occlusion and non-occlusion, determined by the size of the overlapped sub-region relative to the total occlusion ROI. Different tracking methods are then applied for each vehicle based on the severity of occlusion. Trace prediction, using motion features such as velocity and acceleration from a previous frame is applied to cases of severe occlusion. Dictionary and l_2 regularized collaborative representation is applied to partially occluded and non-occluded vehicles. Velazquez-Pupo et al [131] correlated vehicle width and lane width to identify partial

occlusion ROIs in static-camera traffic monitoring applications. The proposed algorithm works under the assumptions that 1.) a vehicle's width should not be greater than the width of one lane, apart from cases in which large vehicles are completely inside the detection ROI, due to perspective effects and 2.) any single vehicle width cannot be larger than 2 lane widths at any time. Feature extraction is then applied to all ROIs which contradict the above assumptions to identify occluded vehicles.

Zhang et al [132] proposed a method to handle vehicle-vehicle occlusions on multiple levels. Intraframe occlusion is detected using convexity cues to determine a compactness ratio and an estimated interior distance of identified convex shapes. This is used to calculate an interior distance ratio indicating the likelihood of vehicle occlusion in the ROI. A cutting line is then calculated and removed in order to separate occluded ROIs into two or more distinct vehicles. Interframe occlusion handling is then conducted by exploiting motion cues between frames. Motion vector analysis is carried out to identify variations in motion within occlusion ROIs indicating the presence of multiple vehicles. Subtractive Clustering is used to detect and remove pixels at the intersecting edge of occluded vehicles in order to separate multiple vehicles in the ROI. Bi-directional tracking level occlusion reasoning is then implemented in order to identify the presence of severe or full occlusions. Occlusion layer images are created consisting of estimated moving vehicle regions occluded by other vehicles. The locations of vehicles in the occlusion layer are updated frame by frame according to their average motion vector. Position of the vehicles in occlusion

layer images is updated using motion information. Once a vehicle re-emerges from the occlusion and is once again visible, it is removed from the occlusion layer. If a vehicle fails to reemerge within a preset number of frames it is removed from the occlusion layer and it is assumed that it has moved out of frame.

Ghasemi and Safabakhsh [133] detected inter-occluding vehicles in tracking applications by monitoring the intersection of tracked boundary boxes between frames in a similar operation to the merge-split approach used in [71]. A template of each vehicle node is maintained to prevent losing specific vehicles within the occluded region. This is carried out by correlating each vehicle node with the overall region to maintain the location of each individual vehicle up to a certain threshold of occlusion, after which a Kalman Filter is used to track vehicle nodes. Fang et al [134] and Huang et al [135] used a feature tracking algorithm to identify and track the visible corner features of each vehicle in a merged occlusion region to prevent loss of specifically tracked vehicles under partial occlusion. Zhang et al [136] proposed a part matching algorithm to track parts of occluded targets between frames. Galceran et al [137] used a hybrid Gaussian Mixture Model (hGMM) to capture multiple hypotheses while tracking vehicles through prolonged occlusions. When the tracked vehicle re-emerges, sensor observations are matched with the estimated occluded states in terms of Kullback-Leibler Divergence (KLD) to associate the tracking data. Min et al [138] used a Support Vector Machine (SVM) combined with Local Binary Pattern (LBP) features in addition to a Convolutional Neural Network (CNN) to reduce the

impact of occlusion during multiple vehicle tracking.

Pham and Lee [139] focused on the appearance of a vehicle's windscreen in order to identify multiple occluding vehicles. Edge detection reasoning is used to identify potential vehicle windscreens and the expected trapezoidal shape characteristics are confirmed using a Hough Transform. A HOG-SVM based classifier is then used to classify vehicles. The classifier is trained using a dataset including images of occluded vehicles for identification of partially occluded vehicles in dense traffic. Ohn-Bar et al [140] merged monocular and stereo vision systems to enhance an appearance based vehicle detection model with depth and motion cues in order to improve detection of partially occluded vehicles.

Wang et al [141] calculated the scores of local visual cues using a trained model to detect semantic parts of partially occluded objects. The spatial relationship of supporting visual concepts are accumulated to infer the existence of known semantic parts, such as wheels, in order to identify the presence of an occluded vehicle. Li et al [142] combined a Region-based Convolutional Neural Network (R-CNN) with a deformable parts based model (DPM) to improve the detection of multiple occluding objects. Many similar research strategies exploit deformable parts based models (DPM) to reduce the impact of occlusion in vehicle detection algorithms [143–146].

Xiang et al [147] trained an occlusion specific detector with 2D images and corresponding 3D Voxel Patterns developed from CAD models, to explicitly itemize occlusion in each image. Proposals of 3D Voxel Patterns during testing allow the

inference of depth in 2D images highlighting occlusion regions of interest. Li et al [148] presented a discriminative AND-OR structure to model occlusions. A synthetically generated CAD dataset representing a wide array of occlusion configurations was used to train a latent structural SVM. Li et al [149] and Wu et al [150] expanded upon this research to propose methods for learning AND-OR models to represent context and occlusion configurations for vehicle detection and viewpoint estimation. Other discriminatively trained methods such as [151, 152] used CNNs to recognise vehicles under different occlusions without the integration of explicit occlusion handling.

Ren et al [100] presented a single stage detection algorithm using a Recurrent Rolling Convolution (RRC) architecture in which each iteration gathers and aggregates relevant features for detection. In this process contextual information can be selectively introduced to the bounding box regressor when required to boost classification performance. This approach can be correlated with the findings the human psychophysical study conducted by Meng and Potter [26], referenced in Section 2.3 of this document which demonstrates that the integration of contextual information can improve the human visual systems ability to tolerate noise such as occlusion. The RRC algorithm has demonstrated consistently high performance on the KITTI Vision Benchmark for the Pedestrian, Cyclist and Vehicle detection test datasets, Table 2.4.

Table 2.4: KITTI High Performing Vehicle Detection Algorithms

Performance (Average Precision)			
Algorithm	Moderate (Partial Occlusion)	Hard (Heavy Occlusion)	Detection Model / Strategy
TuSimple [98, 99]	90.33%	82.86%	<i>CNN</i> <i>Scale-Dependent Pooling,</i> <i>Layer-Wise Cascaded Rejection Classifiers</i>
RRC [7, 100]	90.22%	87.44%	<i>CNN</i> <i>Recurrent Rolling Convolution (RRC)</i>
Deep MANTA [7, 153]	90.03%	80.62%	<i>CNN</i> <i>3D Dimension Estimation, 2D/3D Point Matching</i>
SenseKITTI [154]	90.00%	81.83%	<i>CNN</i> <i>Cascaded Region-Proposal-Network + Fast RCNN</i>
F-PointNet [86]	90.00%	80.80%	<i>Camera Lidar Fusion</i> <i>RGB-D Data, CNN, 2D to 3D Image Extrusion</i>
SINet+ [21]	89.73%	77.82%	<i>CNN</i> <i>Scale Insensitive, Context aware ROI pooling,</i> <i>Multi-branch Decision Network</i>

2.6.2 Two Wheeled Vehicles and Cyclist Detection

Phan et al [155] proposed a vehicle detection algorithm for motorcycles in crowded scenes. Background subtraction was used to model the area from which vehicles can be detected. Geometrical characteristics and the features of object shapes are then assessed using a trained decision tree to identify overlapping blobs of vehicles. Detected occlusion ROIs undergo an iterative segmentation process informed by the analysis of the convex and concave features within the occlusion ROI. Vehi-

cle classification is reattempted after each segmentation cycle until both occluding and occluded vehicles are identified. Other methods of identifying partially occluded cyclists and motorcyclists include helmet or head detection [156–160], parts based models [141, 161, 162] and Vehicle to X (V2X) communication [163–165].

Cai et al [15] proposed a Multi-Scale CNN (MS-CNN) algorithm to increase the detection rate of target objects which are very close to, or further away from the ego vehicle. MS-CNN adapts a feed forward neural network based on object size by exploiting feature maps of several resolutions to detect objects at different scales. Detection is performed at various intermediate network layers whose receptive fields match specific object scales and are then combined to produce multiple scale detection. This multi-scale approach displays a high level of occlusion handling on both the Cyclist and Pedestrian datasets on the KITTI Vision Benchmark. A similar principle was applied in Yang et al [98], the research behind the high ranking “TuSimple” algorithm which achieves state-of-the-art performance on the Pedestrian, Cyclist and Vehicle datasets of the KITTI Vision Benchmark. TuSimple modifies its CNN algorithm based on the size of the ROI using Scale-Dependent Pooling (SDP), then uses a Cascaded Rejection Classifier (CRC) to eliminate negative object proposals. The model is trained using Deep Residual Learning method proposed in [99].

Li et al [114] presented the Tsinghua-Daimler Cyclist Benchmark for identifying occluded cyclists with three degrees of difficulty:

Table 2.5: KITTI High Performing Cyclist Detection Algorithms

Performance (Average Precision)			
Algorithm	Moderate (Partial Occlusion)	Hard (Heavy Occlusion)	Detection Model / Strategy
RRC [7, 100]	76.47%	65.46%	<i>CNN</i> <i>Recurrent Rolling Convolution (RRC)</i>
MS-CNN [15]	74.45%	64.91%	<i>CNN</i> <i>Adaptive CNN based on object scale</i>
TuSimple [98, 99]	74.26%	64.88%	<i>CNN</i> <i>Scale-Dependent Pooling,</i> <i>Layer-Wise Cascaded Rejection Classifiers</i>
Deep3DBox [166]	73.48%	64.11%	<i>CNN</i> <i>3D Pose estimation from 2D bounding box</i>
SDP+RPN [98, 167]	73.08%	64.88%	<i>CNN</i> <i>Scale-Dependent Pooling, Region Proposal Networks</i>
SenseKITTI [154]	72.50%	64.00%	<i>CNN</i> <i>Cascaded Region-Proposal-Network + Fast RCNN</i>

- Easy - Bounding boxes higher than 60 pixels and fully visible.
- Moderate – bounding boxes higher than 45 pixels and < 40% occlusion.
- Hard – cyclists with bounding boxes higher than 30 pixels and < 80% occlusion.

The KITTI [7], KAIST [102] and Multi Object Tracking (MOT) [103] datasets also contain occlusion-specific annotation of cyclists. Table 2.5 provides an overview of high performing cyclist detection algorithms on the KITTI benchmark.

A small number of research projects focus on the issue of e-scooter rider detection

[168–171]. However no known research has been carried out to date on the detection and classification of e-scooter riders under partial occlusion.

2.7 Occluded Objects and Traffic Signs

In order to be effective in culturally diverse real-world applications, semi-autonomous and autonomous vehicles must be able to detect and classify both known and unknown objects. The following section provides an overview of occlusion handling strategies used in the identification and tracking of generic and specific objects.

Hoiem et al [39] outlined an iterative segmentation process for identifying occlusion boundaries from a single image. This method carries out initial segmentation, then uses multiple occlusion cues to estimate a soft boundary map using a Conditional Random Field (CRF) model which in turn is used to produce a refined segmentation output. This process is repeated several times to further refine the occlusion boundaries by reusing the segmentation output as an input to increase confidence and remove weak boundaries.

Chu and Krzyzak [172] evaluated the performance of Support Vector Machines (SVM), Convolutional Neural Networks (CNN) and Deep Belief Networks (DBN) when tasked with identifying partially occluded objects. Each of the models were consistently trained on datasets of non-occluded, occluded and mixed (both occluded and non-occluded) objects and their performance in detecting partially occluded objects was compared. These experiments demonstrated that training a model exclu-

sively on a non-occluded dataset leads to poor results when identifying occluded objects. The authors also found that training models on a mixed dataset of both occluded and non-occluded objects yielded the best results overall, and that when trained exclusively on occluded objects, DBNs can still provide a high level of accuracy (up to 69%) when identifying non-occluded objects. The authors concluded that generative models such as DBNs do not exceed the performance of purely discriminative models such as CNN and SVM when tasked with identifying partially occluded objects. These findings appear to contradict the work of Susskind et al [173] which demonstrated increased performance by exploiting the generative ability of DBN in facial recognition under partial occlusion. However, Chu and Krzyzak [172] indicate that such differences may be primarily due to the implementation of additional reconstructive processes rather than the architecture or training method used.

Cavagna et al [174] proposed a Spatiotemporal Reconstruction Tracking Algorithm (SpARTA) to track featureless objects through occlusion. This approach represents each target as a cloud of 3D points. When occlusion occurs, represented by a 3D cluster or multiple partially-separated dense point clouds, the algorithm separates ambiguous connected components into partitions corresponding to the trajectory of a single target by defining and solving an optimisation problem. A cost function is developed to partition trajectories using attractive links connecting points that are close enough to be related and repulsive links to separate parts over a des-

ignated spatio-temporal threshold. Liu et al [175] presented an occlusion robust traffic sign classifier based on extended sparse representation classification (ESRC). In addition to a content dictionary of known traffic signs, this method includes an occlusion dictionary to represent common occlusion cases of different signs to increase recognition under partial occlusion. Huang et al [176] proposed a method of detecting and analysing the degree of traffic sign occlusion using mobile laser scanner point clouds. Other strategies for traffic sign recognition under partial occlusion use adaboost-selected Haar-like features and SVM [177], HOG and SVM [178], fuzzy shape recognition [179] and 3D-reconstruction from multiple views [180].

2.8 Conclusions and Remaining Challenges

Humans have a natural ability to detect, recognise and track partially occluded objects. Many object detection strategies can be correlated back to psychophysical studies of the human visual system referenced in Section 2.3. Although many of the current state-of-the-art methods for partially occluded object detection are primarily CNN based, the HVS uses prior knowledge as a guide only, also using real time assessment of the contextual and visual characteristics of a scene to inform decision making. Humans exploit a hierarchical and cooperative system for the identification of objects. Information is relayed from the retina to the visual cortex where initial processing and identification of simple objects is carried out. For more complex scenes such as in the case of partial occlusion, communication occurs between the

visual cortex and the prefrontal cortex where memory and learning are used to help identify the object [28]. Despite recent breakthroughs in CNN based algorithms, traditional computer vision based detection methods remain as relevant as ever as the performance increases offered by stand-alone deep learning algorithms for object detection in the automotive environment have begun to plateau. The convergence of deep learning and traditional occlusion handling methods, as well as further insight on bioinspired occlusion handling, has the potential to inform more robust object detection algorithms. This is reflected in the consistency of detection methods which replicate bioinspired occlusion handling by combining visual occlusion cues and prior contextual knowledge. Such methods utilise a multibranch approach based on the scale of the target ROI or the level of occlusion, whether the target is under no occlusion, partial occlusion or heavy occlusion.

The deformable nature and smaller scale of vulnerable road users such as pedestrians and cyclists present an added complexity to the occlusion challenge. Current high performing vehicle detection algorithms are commonly able to detect approximately 90% of Partial Occlusions and approximately 80% of Heavy Occlusions according to the KITTI Vision Benchmark Suite Leader board [7]. A considerable amount of work has yet to be carried out on Pedestrian and Cyclist detection which is still only in the region of 65%-75% detectable under partial and heavy occlusions. A significant knowledge gap exists for the detection of e-mobility users such as e-scooter riders under partial occlusion.

In order to achieve the performance required for safe autonomous driving, an algorithm or set of algorithms, must consistently generalise to reach state of the art performance in all benchmarks, cultures and environmental conditions. Additionally, and perhaps most challengingly, any successful approach must also have the computational efficiency to robustly identify objects in real time. The process of accurately assessing algorithm performance for the detection of partially occluded objects is a difficult one. There are a wide variety of test datasets available for object detection based on desired target, environmental conditions and sensing methods. Inconsistency between each dataset’s definition and annotation level of occluded targets and the metrics used, present difficulties when attempting to accurately quantify performance. Although comparative datasets provide indicative performance of new algorithms, no definitive set of metrics exists to ensure algorithm performance is reported in an objective manner. Taking the KITTI Vision Benchmark and the Caltech Pedestrian Dataset for example, both datasets can be used to compare performance for partially occluded pedestrians as shown in Table 2.1 and Table 2.2 respectively. However, each benchmark varies greatly in the definition of occlusion and the annotation methods used to apply occlusion labels. Table 2.3 shows the definition of occlusion labels used in popular detection benchmarks. The KITTI Vision Benchmark defines “Heavy Occlusion” as any pedestrian instance that is perceived by the human annotator to be “difficult to see” whereas the Caltech Pedestrian Dataset specifies “Heavy Occlusion” as pedestrians that are between “35%-80% Occluded”.

In addition, each benchmark uses different but highly subjective methods for applying the occlusion labels, Section 2.5.7, leading to inconsistent reporting of algorithm performance between benchmarks. A knowledge gap exists for a robust, repeatable method of occlusion level classification which provides objective, fine-grained occlusion level annotation for pedestrian detection benchmarks.

Chapter 3 focuses on addressing this knowledge gap through the development of an objective metric and occlusion level annotation method for the occurrence and severity of occlusion in an image sequence.

Chapter 3

An Objective Method for Pedestrian Occlusion Level Classification

3.1 Summary

Pedestrian detection is among the most safety-critical features of driver assistance systems for autonomous vehicles. One of the most complex detection challenges is that of partial occlusion, where a target object is only partially available to the sensor due to obstruction by another foreground object. A number of current pedestrian detection benchmarks provide annotation for partial occlusion to assess algorithm performance in these scenarios, however each benchmark varies greatly

in their definition of the occurrence and severity of occlusion. In addition, current occlusion level annotation methods commonly contain a high degree of subjectivity by the human annotator. This can lead to inaccurate or inconsistent reporting of an algorithm’s detection performance for partially occluded pedestrians, depending on which benchmark is used. This chapter presents a novel, objective method for pedestrian occlusion level classification for ground truth annotation. Occlusion level classification is achieved through the identification of visible pedestrian keypoints and through the use of a novel, effective method of 2D body surface area estimation. Experimental results demonstrate that the proposed method more accurately reflects the pixel-wise occlusion level of pedestrians than the current state of the art and is effective for all forms of occlusion, including challenging edge cases such as self-occlusion, truncation and inter-occluding pedestrians.

3.2 Introduction

Robust pedestrian detection is one of the most safety-critical features of driver assistance systems and autonomous vehicles. Pedestrian detection is particularly challenging due to the deformable nature and irregular profile of the human body in motion and the inconsistency of color information due to clothing, that can enhance or camouflage any part of a pedestrian. Pedestrian detection systems have improved significantly in recent years with the proliferation of deep learning based solutions and the availability of larger and more diverse datasets. Despite this, many challenges still

exist before the detection capabilities required for safe autonomous driving is reached. One of the most complex scenarios is that of partial occlusion, where a target object is only partially available to the sensor due to obstruction by another foreground object. The frequency and variety of occlusion in the automotive environment is substantial and is impacted by both natural and man-made infrastructure as well as the presence of other road users [181–183]. Pedestrians can be occluded by static or dynamic objects, may inter-occlude (occlude one another) such as in crowds, and self-occlude - where parts of a pedestrian overlap. State of the art pedestrian detection solutions claim a detection performance of approximately 65%-75% of partially and heavily occluded pedestrians respectively using current benchmarks [10, 184–186]. However, the definition of the occurrence and severity of occlusion varies greatly, and a high degree of subjectivity is used to categorise pedestrian occlusion level in each benchmark as shown in Table 2.3. In addition, occurrences of self occlusion, where one part of the body occludes another, has typically been overlooked entirely when categorizing occlusion level. This can lead to inaccurate or inconsistent reporting of a pedestrian detection algorithm’s performance, depending on which dataset is used to verify detection performance [10, 187]. In order to address this issue, a universal metric and an objective, repeatable method of occlusion level classification is required for ground truth annotation so that algorithms can be evaluated and compared on an equal scale.

This research proposes a novel, objective and consistent method for pedestrian

occlusion level classification for ground truth annotation of partially occluded pedestrians. The proposed method more accurately represents the pixel-wise occlusion level than the current state of the art and works for all forms of occlusion including challenging edge cases such as self-occlusion, inter-occluding pedestrians and truncation.

The contributions of this research are threefold: 1. A novel, objective method for pedestrian occlusion level classification for ground truth annotation is presented. 2. A novel method for estimating the visible 2D body surface area of pedestrians in images. 3. The proposed method is the first occlusion level classifier to infer the level of pedestrian self-occlusion.

3.3 Related Work

Section 2.5.7 provides an overview of current occlusion level classification methods for pedestrian detection, pedestrian analysis for flood level assessment and commonly used methods for estimating total body surface area.

3.4 Methodology

An objective method for occlusion level classification is proposed, which removes the subjectivity of the human annotator and more accurately reflects the pixel wise occlusion level than the current state of the art [5, 7–9, 53, 105, 114]. Occlusion level

classification consists of the following steps: 1. Keypoint detection is applied to the input image in order to identify the presence and visibility of specific semantic parts for each pedestrian instance. 2. A visibility threshold is applied to identify occluded keypoints. 3. MaskRCNN is applied to define the pedestrian mask area and results are cross-referenced with detected keypoints to confirm which keypoints are occluded within the image. 4. Visible keypoints are then grouped into larger semantic parts and the total visible surface area is calculated using the 2D body surface area estimation method outlined in Section 3.4.2 and Figure 3.3. The proposed method classifies occlusion level for all forms of pedestrian occlusion, including challenging edge cases such as self occlusion, inter-occluding pedestrians and truncation. An overview of the classification pipeline is shown in Figure 3.1 and qualitative examples of the classifier output for multiple scenarios can be seen in Figure 3.4.

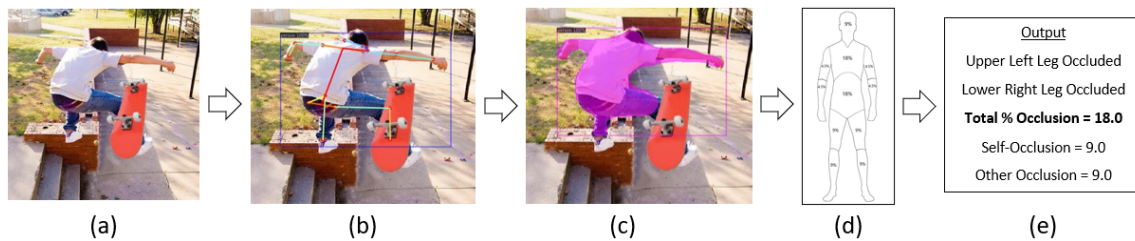


Figure 3.1: Occlusion level classification overview. (a) Input image (b) Apply keypoint detection to each pedestrian instance and assess keypoint visibility to identify occluded keypoints (c) Correlate visible keypoints with the pedestrian mask to confirm visibility and occlusion type (d) Calculate total visible surface area (e) Output occlusion level classification.

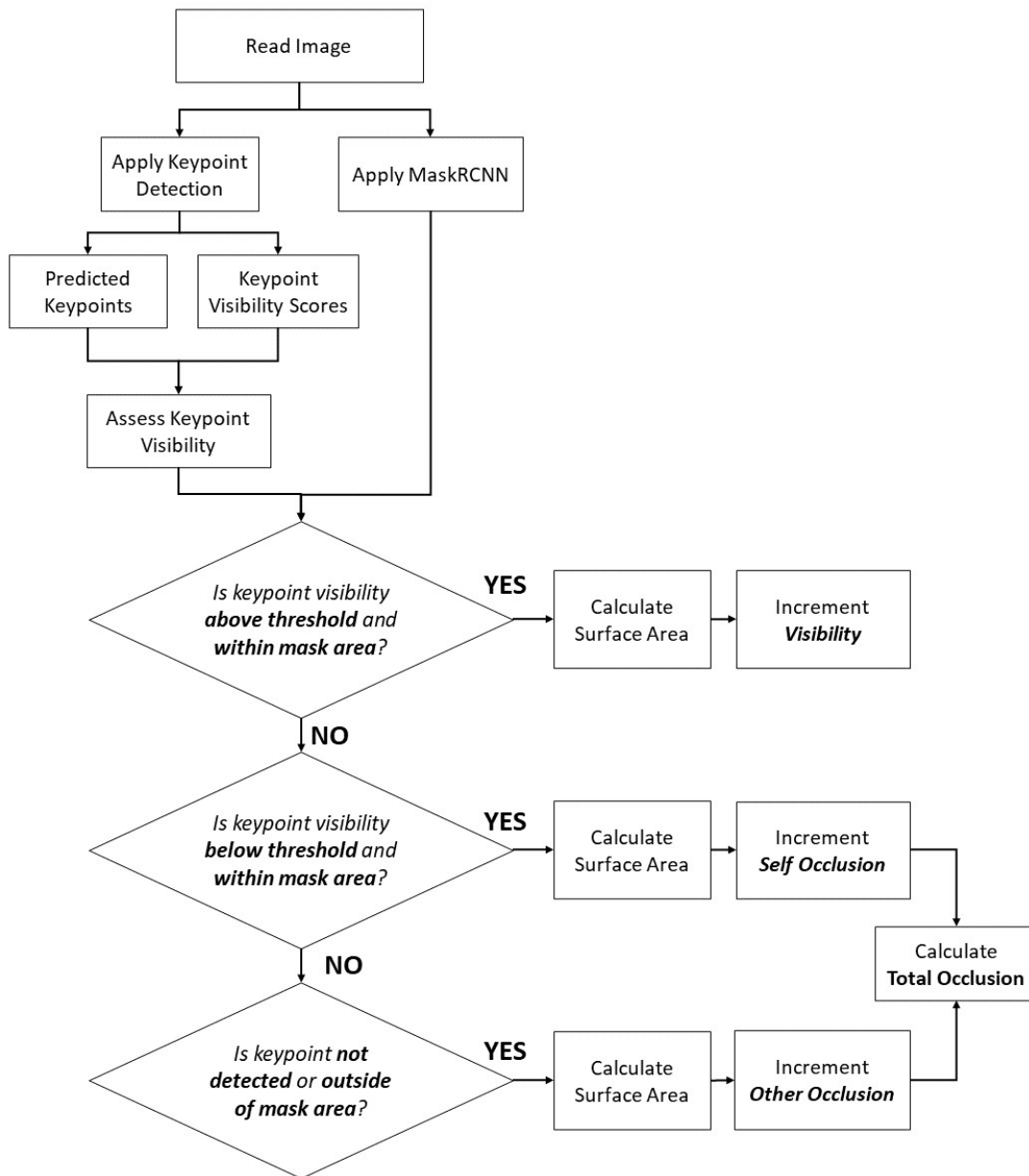


Figure 3.2: Occlusion Level Classification Pipeline.

3.4.1 Occluded Keypoint Detection

Keypoint detection is carried out by a Faster RCNN based keypoint detector using pretrained weights from Detectron2 [188]. The model uses a ResNet-50-FPN backbone and is trained using the COCO keypoints dataset [189]. The keypoint detector outputs 17 keypoints on the human body in addition to a visibility score for each predicted keypoint. Predicted keypoints include shoulders, elbows, wrists, hips, knees and ankles as well as facial characteristics such as nose, eyes and ears. A two-step process is then applied to determine the visibility of keypoints in an image. First, a threshold is applied to the keypoint visibility score returned from the keypoint detector. The coordinates of each visible keypoint are then cross-referenced with the pedestrian mask generated by MaskRCNN [190] to confirm the keypoint location is within the pedestrian mask region in the image. This two-step process increases the identification of occluded keypoints in complex cases such as self-occlusion and inter-occluding pedestrians where the keypoint visibility score is low however the estimated keypoint location may be masked due to the occluding pedestrian region. The presence of specific grouped keypoints indicates the presence of semantic body parts as outlined in Table 3.1.

3.4.2 2D Body Surface Area Estimation

The “Wallace Rule of Nines” [125] is a time-tested method for determining total body surface area of the average adult. Although effective in the assessment of

the body surface area of physical pedestrians, the Rule of Nines is not suitable for assessing the visible surface area of pedestrians in 2D images due to the 3D nature of the human body. An adapted version of the Rule of Nines is proposed for use in determining the visible body surface area of 2D pedestrian images for occlusion level classification. The original proportions of the Rule of Nines have been adjusted respectively to compensate for only one side of the body being visible at any one time, as in the case of 2D images. The proposed method for 2D body surface area estimation is shown in Figure 3.3. Detected keypoints are related to the semantic body areas using the lookup table shown in Table 3.1. Examples of the classification output are shown in Figure 3.4.

3.5 Validation

Qualitative Validation was carried out by applying the proposed method to a wide range of images containing various pedestrian poses, backgrounds and multiple forms of occlusion, including cases of self-occlusion, inter-occluding pedestrians, and truncation. Occlusion level and the occluded semantic parts of each pedestrian instance was deduced using the proposed occlusion level classification method. Human visual inspection was then used to verify the performance of the occlusion level classifier in each case. A custom dataset of 320 images, compiled from multiple publicly available sources including [8, 9, 191, 192], was used in this validation step to ensure a wide diversity of pedestrian occlusion scenarios. Examples of the qualitative validation

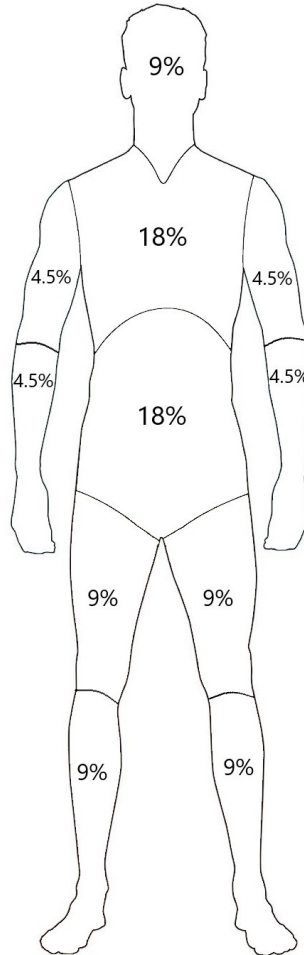

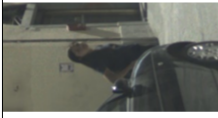




Figure 3.3: 2D Body Surface Area.

are provided in Figure 3.4.

3.5.1 Quantitative Validation

Quantitative validation was carried out by comparing the proposed method with the calculated pixel-wise occlusion level, derived using MaskRCNN [190], and the

Occluded Person					
Proposed Method	Total Occlusion	9%	45%	67.5%	9%
	Occluded Parts	Left Arm	Left Arm, Left Leg, Right Leg	Left Arm, Lower Right Arm, Bottom Torso, Left Leg, Right Leg	Lower Right Leg
<i>State of the Art</i>	Self-Occlusion	0%	9%	0%	0%
	Other Occlusion	9%	36%	67.5%	54%
		<i>Partial</i>	<i>Heavy</i>	<i>Heavy</i>	<i>Partial</i>
		<i>City/Persons</i>			





Occluded Person					
Proposed Method	Total Occlusion	45%	36%	63%	58.5%
	Occluded Parts	Top Torso, Right Arm, Right Leg	Left Leg, Right Leg	Right Arm, Bottom Torso, Left Leg, Right Leg	Top Torso, Bottom Torso, Upper Left Arm, Left Leg
<i>State of the Art</i>	Self-Occlusion	27%	0%	0%	27%
	Other Occlusion	18%	36%	63%	58.5%
		<i>N/A</i>	<i>Heavy</i>	<i>Heavy</i>	<i>N/A</i>
		<i>City/Persons</i>			

Figure 3.4: Qualitative validation results. Occlusion level is displayed below each image using the proposed method and the current state of the art as described in CityPersons [8]. Examples are shown for cases of inter-class occlusion, self occlusion and inter-occluding pedestrians. Images containing multiple pedestrian instances read from left to right. All images are compiled from publicly available sources.

Table 3.1: Percentage of visible body surface area (BSA) and related keypoints for each semantic body part.

Body Part (% BSA)	Related Keypoints
Head (9%)	Nose or Eyes or Ears
Upper Torso (18%)	Left Shoulder and Right Shoulder
Upper Left Arm (4.5%)	Left Shoulder and Left Elbow
Lower Left Arm (4.5%)	Left Elbow and Left Wrist
Upper Right Arm (4.5%)	Right Shoulder and Right Elbow
Lower Right Arm (4.5%)	Right Elbow and Right Wrist
Lower Torso (18%)	Left Hip and Right Hip
Upper Left Leg (9%)	Left Hip and Left Knee
Lower Left Leg (9%)	Left Knee and Left Ankle
Upper Right Leg (9%)	Right Hip and Right Knee
Lower Right Leg (9%)	Right Knee and Right Ankle

current state of the art as described in CityPersons [8] for both visible and progressively occluded pedestrians. In order to determine the pixel-wise occlusion, the total pixel area must be calculated for both the fully visible pedestrian and the same pedestrian under occlusion. To achieve this, a custom dataset of 200 images was created, including a wide range of occlusion scenarios and challenging pedestrian poses such as walking, running and cycling. MaskRCNN [190] was applied to a fully visible reference image and the masked pixel area ($MaskArea_{full}$) was calculated for each pedestrian instance. Occlusions were then superimposed on the reference image and



Figure 3.5: Quantitative validation dataset sample images. The custom dataset consists of 200 images covering a wide range of pedestrian poses and superimposed occlusions designed to test extreme occlusion cases from 0% to 99% occluded. The Amodal Human Perception Test Dataset [193] contains 56 images across multiple domains. All images are compiled from publicly available sources.

the remaining visible pedestrian pixel area ($MaskArea_{occ}$) is calculated in order to determine the pixel-wise occlusion ratio, Equation 3.1.

$$Occ_{pixel} = \frac{MaskArea_{occ}}{MaskArea_{full}} \quad (3.1)$$

The proposed method was then compared with the pixel-wise occlusion level and the method described in CityPersons [8] to determine the pixel-wise accuracy of the proposed occlusion level classifier. More subjective occlusion level classification methods such as those used in [5,7,9,105] are omitted for the purposes of this testing. Quantitative validation results on the custom dataset are provided in Figure 3.6.

Further validation is carried out on the Amodal Human Perception (AHP) test dataset [193] to assess classification performance on an independent multi-domain dataset. The AHP dataset contains 56 images of partially occluded persons across a wide range of activities and poses as well as the modal and amodal mask for each pedestrian instance. Quantitative validation results on the AHP dataset are provided in Figure 3.7.

3.6 Discussion and Analysis

An objective method for occlusion level classification is proposed. The qualitative validation results shown in Figure 3.4 demonstrate the capability of the proposed method for classifying occlusion level for all forms of occlusion, including challenging edge cases such as self-occlusion, truncation, and inter-occluding pedestrians. By

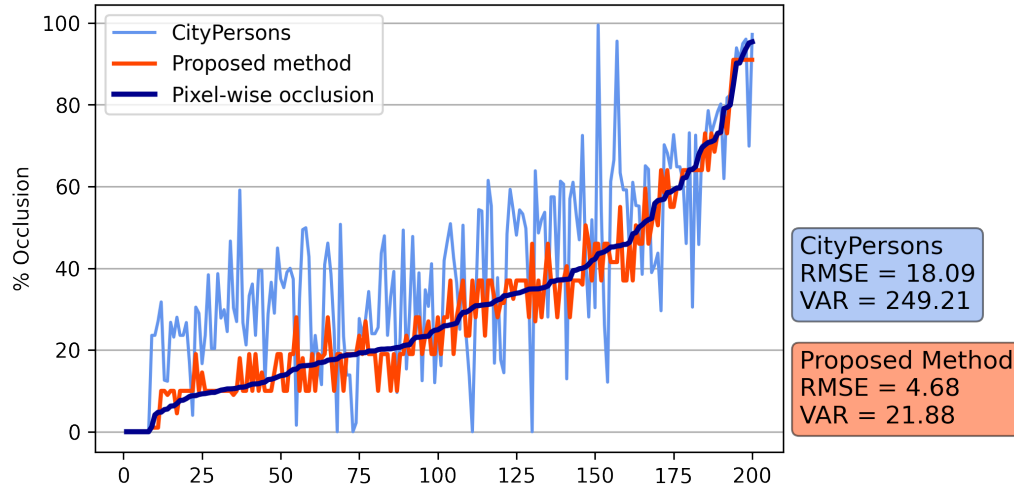


Figure 3.6: Quantitative Evaluation Results 1. The proposed method is compared with the pixel-wise occlusion level as produced by MaskRCNN [190] and the current state of the art as described in CityPersons [8] for a dataset of 200 images, designed to test extreme occlusion cases from 0%-99% occluded. Results demonstrate that the proposed method (RMSE=4.68) is a significant improvement over the state of the art (RMSE=18.09) when plotted against the pixel-wise occlusion level.

removing the subjectivity of a human annotator, the proposed method is more robust and repeatable than the current state of the art and is suitable for the objective comparison of pedestrian detection algorithms, regardless of the benchmark used. Classification of pedestrian self-occlusion, heretofore ignored in the assessment of partially occluded pedestrians, may have a large impact on assessing the detectability of pedestrians using modern techniques. This is especially relevant in scenarios where detection confidence is linked to the presence of key salient features which may be self-occluded by the target pedestrian in the image. More detailed analysis of detection performance in cases of self-occlusion will increase our understanding of

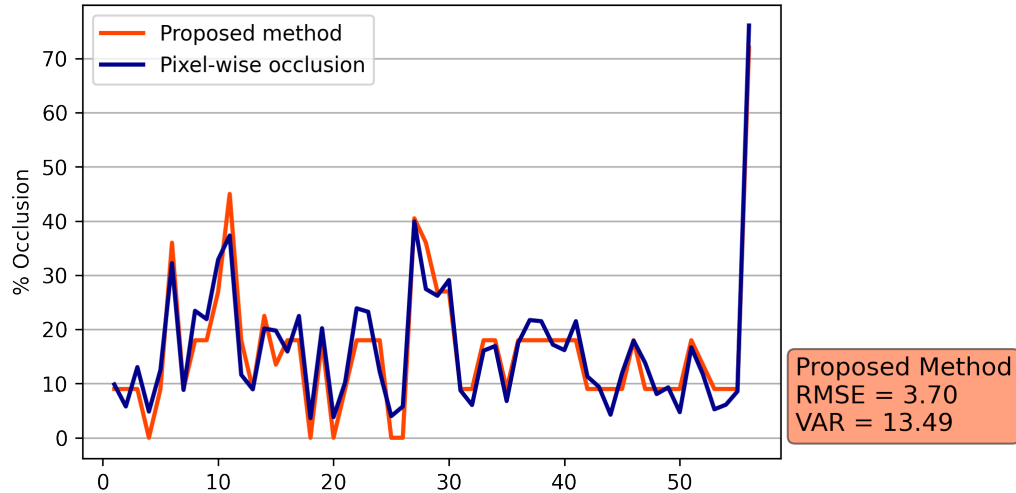


Figure 3.7: Quantitative Evaluation Results 2. The proposed method is compared with the pixel-wise occlusion level for the Amodal Human Perception Test Dataset [193]. Results demonstrate that the proposed method closely reflects the pixel-wise occlusion level for challenging images across multiple domains.

the behaviour of deep learning-based detection routines. Characterisation of detection performance for what were previously considered “visible” pedestrians, in cases where the algorithm specific informative value of a pedestrian is occluded can help to identify potential failure modes of current state of the art pedestrian detection systems.

The quantitative validation results shown in Figure 3.6 and Figure 3.7, demonstrate the proposed method’s capability in representing the “real world” or pixel-wise occlusion value for challenging pedestrian poses, regardless of the severity or form of occlusion. The proposed method of 2D body surface area estimation shown in

Figure 3.3, derived from the “Wallace Rule of Nines”, has proven effective in calculating the visible area of partially occluded pedestrians for a wide range of pedestrian poses and occlusion scenarios. Further analysis of the quantitative validation results clearly displays an improvement over the current state of the art [8] when compared to the pixel-wise occlusion value.

3.6.1 Challenging Image Frames

Figure 3.8 provides a sample of the classifier performance for challenging detection scenarios as well as highlighting classification errors that can occur for indistinct pedestrian instances in particular frames. Missed detections or false negatives can occur as a result of low detection confidence of the keypoint detector or MaskRCNN due to excessive motion blur, camera artifacts or low images resolution. Detection confidence is reduced in scenarios where the pedestrian outline closely matches that of the image background. Figure 3.8 (a), (b) and (c) successfully classify pedestrian occlusion level in cases of heavy occlusion, image glare and low resolution respectively, demonstrating that the occlusion level annotation method is not impacted by image quality once the pedestrian instance has been accurately detected. In each case, the pedestrian outline distinctly differs from the image background. In similar scenarios where the pedestrian outline and the image background are less diverse, such as in Figure 3.8 (h), (j) and (k), detection confidence is reduced resulting in a false negative. Keypoint errors can occur in complex detection scenarios which can result

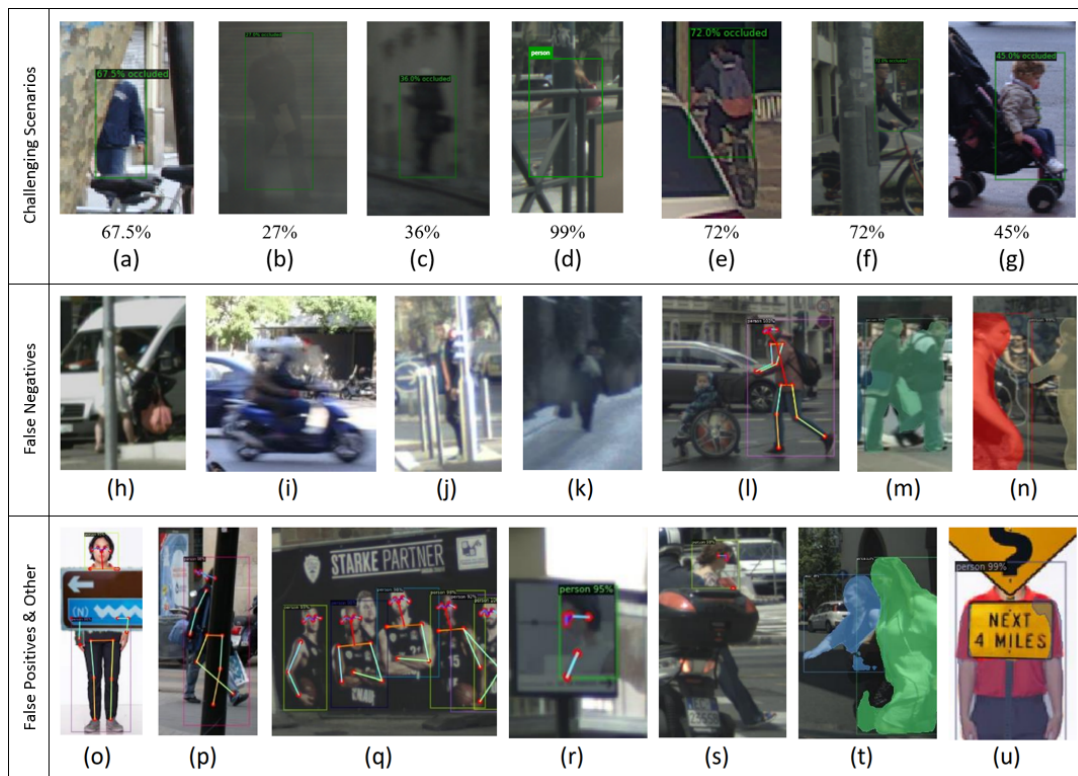


Figure 3.8: Examples of challenging image frames, false negatives and false positives. The top row provides examples of challenging detection scenarios and displays the occlusion level below each image. The middle row provides examples of false negatives and the bottom row provides examples of false positives and other classification errors.

in incorrect classification for a particular frame. Occurrences of these have been noted in cases where a pedestrian instance is highly segmented by the occluder, prompting the algorithm to propose multiple pedestrian instances or omitting sections of a pedestrian that appear to be unconnected to the primary pedestrian instance due to intersecting occlusion. Examples of these occurrences can be seen in Figure 3.8

(o), (p) and (s). Similarly, pedestrian mask errors can also occur in particularly challenging frames. Mask errors can include mask leakage, which can falsely indicate the presence of occluded keypoints, Figure 3.8 (u), and incomplete or imprecise masks which can lead to the false omission of specific keypoints or pedestrian instances as shown in Figure 3.8 (m), (n) and (t). Although the proposed method is designed to focus on pedestrians, other road users such as cyclists, motorcyclists and children in strollers may be classified as occluded pedestrians. In addition, person depictions in advertising images and other media may be classified as pedestrians by the algorithm. Many of the misclassification errors presented can be reduced by further improvement in keypoint and pedestrian mask detection models which can be integrated into the detection pipeline as technology progresses.

3.7 Conclusions

This research proposes an objective method of pedestrian occlusion level classification for ground truth annotation. The proposed method uses keypoint detection and mask segmentation to identify and determine the visibility of the semantic parts of partially occluded pedestrians and calculates the percentage occluded body surface area using a novel, effective method for 2D body surface area estimation. The proposed method removes the subjectivity of the human annotator used by the current state of the art, in turn increasing the robustness and repeatability of pedestrian occlusion level classification. Qualitative and quantitative validation demonstrates

the effectiveness of the proposed method for all forms of occlusion including challenging edge cases such as self-occlusion and inter-occluding pedestrians. Experimental results show a significant improvement in Root Mean Squared Error (4.68) and Variance (21.88) over the current state of the art (RMSE = 18.09, VAR = 249.21) when plotted against the pixel-wise pedestrian occlusion level.

Table 2.3 displays the categories of occlusion level annotation for a number of popular pedestrian detection benchmarks. Current benchmarks are inconsistent in their definition of the occurrence and severity of occlusion, use subjective methods for annotation and group occlusion instances into two to three broad categories such as low, partial and heavy occlusion. A knowledge gap remains for the provision of a fine-grained, objective dataset for the benchmarking of detection performance for partially occluded pedestrians.

Chapter 4 progresses the research carried out in Chapter 3 by using the proposed methodology to create an objective test dataset for the characterisation of detection performance for partially occluded pedestrians. Performance characterisation is carried out for a number of popular pedestrian detection routines in order to provide detailed analysis of the impact of partial occlusion on pedestrian detectability.

Chapter 4

The Impact of Partial Occlusion on Pedestrian Detectability

4.1 Summary

Robust detection of vulnerable road users is a safety critical requirement for the deployment of autonomous vehicles in heterogeneous traffic. One of the most complex outstanding challenges is that of partial occlusion where a target object is only partially available to the sensor due to obstruction by another foreground object. A number of leading pedestrian detection benchmarks provide annotation for partial occlusion, however each benchmark varies greatly in their definition of the occurrence and severity of occlusion. Recent research demonstrates that a high degree of subjectivity is used to classify occlusion level in these cases and occlusion

is typically categorised into 2-3 broad categories such as “partially” and “heavily” occluded. In addition, many pedestrian instances are impacted by multiple inhibiting factors which contribute to non-detection such as object scale, distance from camera, lighting variations and adverse weather. This can lead to inaccurate or inconsistent reporting of detection performance for partially occluded pedestrians depending on which benchmark is used. This chapter introduces a novel, objective benchmark for partially occluded pedestrian detection to facilitate the objective characterisation of pedestrian detection models. Characterisation is carried out on seven popular pedestrian detection models for a range of occlusion levels from 0-99%. Results demonstrate that pedestrian detection performance degrades, and the number of false negative detections increase as pedestrian occlusion level increases. Of the seven popular pedestrian detection routines characterised, CenterNet has the greatest overall performance, followed by SSDlite. RetinaNet has the lowest overall detection performance across the range of occlusion levels.

4.2 Introduction

Accurate and robust pedestrian detection systems are an essential requirement for the safe navigation of autonomous vehicles in heterogeneous traffic. Leading pedestrian detection systems claim a detection performance of approximately 65%-75% of partially and heavily occluded pedestrians respectively using current benchmarks [10, 184–186]. However, recent research as described in Chapter 3 [12] demonstrates

that the definition of occurrence and severity of occlusion varies greatly, and a high degree of subjectivity is used to categorise pedestrian occlusion level in each benchmark. Occlusion is typically split into 2-3 broad, loosely defined, categories such as “partially” or “heavily” occluded [7–9]. In addition, many pedestrian instances are impacted by multiple inhibiting factors that contribute to non-detection such as object scale, distance from camera, lighting variations and adverse weather. This makes it difficult to determine if the primary factor for non-detection is the severity of occlusion alone, and can lead to inaccurate or inconsistent reporting of detection performance for partially occluded pedestrians depending on which benchmark is used. A knowledge gap exists for a methodology for objective, detailed occlusion level analysis for pedestrian detection across the complete spectrum of occlusion levels. Use of an objective, fine grained occlusion specific benchmark will result in more reliable, consistent and detailed analysis of pedestrian detection algorithms for partially occluded pedestrians.

This chapter presents a novel, objective benchmark for partially occluded pedestrian detection to facilitate the objective characterisation of pedestrian detection models. Objective characterisation of occluded pedestrian detection performance is carried out for seven popular pedestrian detection routines for a range of occlusion levels from 0-99%. The contributions of this chapter are as follows: 1. A novel, objective, test benchmark for partially occluded pedestrian detection is presented. 2. Objective characterisation of pedestrian detection performance is carried out for

seven popular pedestrian detection routines.

4.3 Related Work

A number of popular pedestrian detection benchmarks provide annotation of pedestrian occlusion level to determine the relative detection performance for partially occluded pedestrians, however, benchmarks can be inconsistent in their definition of the occurrence and severity of occlusion as discussed in Chapter 3. Dollar et al [6] provides analysis on occluded pedestrians based on the Caltech Pedestrian Dataset [5]. Caltech Pedestrian estimates the occlusion ratio of pedestrians by annotating two bounding boxes, one for the visible pedestrian area and one for the annotators' estimate of the total pedestrian area. Pedestrians are categorised into two occlusion categories, "partially occluded", defined as 1-35% occluded and "heavily occluded", defined as 35-80% occluded. Any pedestrians suspected to be more than 80% occluded are labelled as fully occluded. Analysis of the frequency of occlusion on the Caltech Pedestrian Dataset demonstrated that over 70% of pedestrians were occluded in at least one frame, highlighting the frequency of occurrence of pedestrian occlusion in the automotive environment. The Eurocity Persons [9] Dataset categorises pedestrians according to three occlusion levels: low occlusion (10%-40%), moderate occlusion (40%-80%), and strong occlusion (larger than 80%). Classification is carried out by human annotators in a similar manner to the Caltech Pedestrian Dataset. The full extent of the occluded pedestrian is estimated, and the approximate level of

occlusion is then estimated to be within one of the three defined categories. Citypersons [8] calculate occlusion levels by drawing a line from the top of the head to the middle of the two feet of the occluded pedestrian. Human annotators are required to estimate the location of the head and feet if these are not visible. A bounding box is then generated for the estimated full pedestrian area using a fixed aspect ratio of $0.41(\text{width}/\text{height})$. This is then compared to the visible area bounding box to denote occlusion level. These estimates of occlusion level are then categorised into two levels, “reasonable” ($\leq 35\%$ occluded) and “heavy occlusion” (35%-75%). Similar approaches are taken in [38, 113, 117, 118, 194]. The KITTI Vision Benchmark [7] and Multispectral Pedestrian Dataset [105] tasked human annotators with marking each pedestrian bounding box as “visible”, “semi-occluded”, “fully-occluded”.

Although these methods are useful for the relative comparison of detection performance on specific datasets, the occlusion categories used are broad (usually 2 to 3 categories), are inconsistent from benchmark to benchmark, and involve a high degree of subjectivity by the human annotator, Chapter 3 [11, 12]. A knowledge gap exists for a detailed, objective benchmark to compare pedestrian detection performance for partially occluded pedestrians in a more repeatable and robust manner. Many pedestrian detection analysis papers [6, 37, 38, 187, 195–200] and occlusion-specific survey papers [10, 184, 186, 201, 202] highlight the outstanding challenges posed by occluded pedestrians, however, no known objective characterisation of pedestrian detection performance spanning the spectrum of occlusion levels has been carried out to date.

Chapter 3 [12] proposes an objective method of occlusion level annotation and visible body surface area estimation of partially occluded pedestrians. Keypoint detection is applied to identify semantic body parts and findings are cross-referenced with a visibility score and the pedestrian mask in order to confirm the presence or occlusion of each semantic part. A novel method of 2D body surface area estimation based on the “Wallace rule of Nines” [11, 125] is then used to quantify the total occlusion level of pedestrians.

4.4 Methodology

A novel occluded pedestrian test dataset, containing 820 person instances in 724 images, has been created in order to characterise pedestrian detection performance across a range of occlusion levels from 0-99%. A diverse mix of images are used ensure that a wide variety of target pedestrians, pedestrian poses, backgrounds, and occluding objects are represented. The dataset is sourced from three main categories of images: 1) The “occluded body” subset of the partial re-identification dataset “Partial ReID” provided by Zheng *et al* [203], 2) The Partial ReID “whole body” subset [203] with custom superimposed occlusions and 3) Images collated from publicly available sources including [8, 9, 11, 191]. All images are annotated using the objective occlusion level classification method described in Chapter 3 [12]. Complex cases at very high occlusion rates were manually verified using the method of 2D body surface area estimation presented in Chapter 3 [12]. Each occlusion level contains

a minimum of 55 pedestrian instances. Dataset statistics by occlusion level and a sample of the test dataset can be seen in Figure 4.1 and Figure 4.2 respectively.

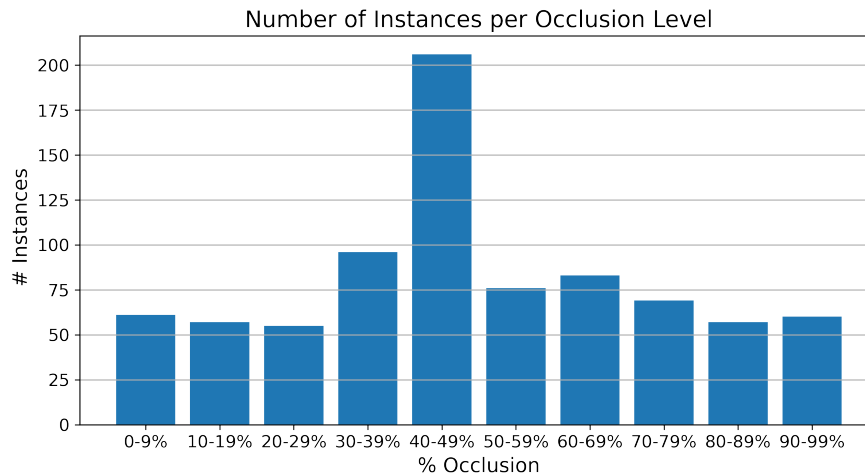


Figure 4.1: Dataset Statistics. The number of pedestrian instances per occlusion level. The custom dataset contains 820 pedestrian instances under progressive levels of occlusion from 0-99%.

4.4.1 Pedestrian Detection Models

Performance characterisation was carried out on seven popular pedestrian detection models. All models use publicly available pretrained weights from two popular model zoos [204,205] and are trained using the COCO “train 2017” dataset [189]. An overview of the pedestrian detection models and their performance on the proposed dataset can be seen in Table 4.1.

The pedestrian detection models chosen for characterisation can be divided into 3 categories: Two-Stage Frameworks, One-Stage Frameworks and Keypoint Esti-

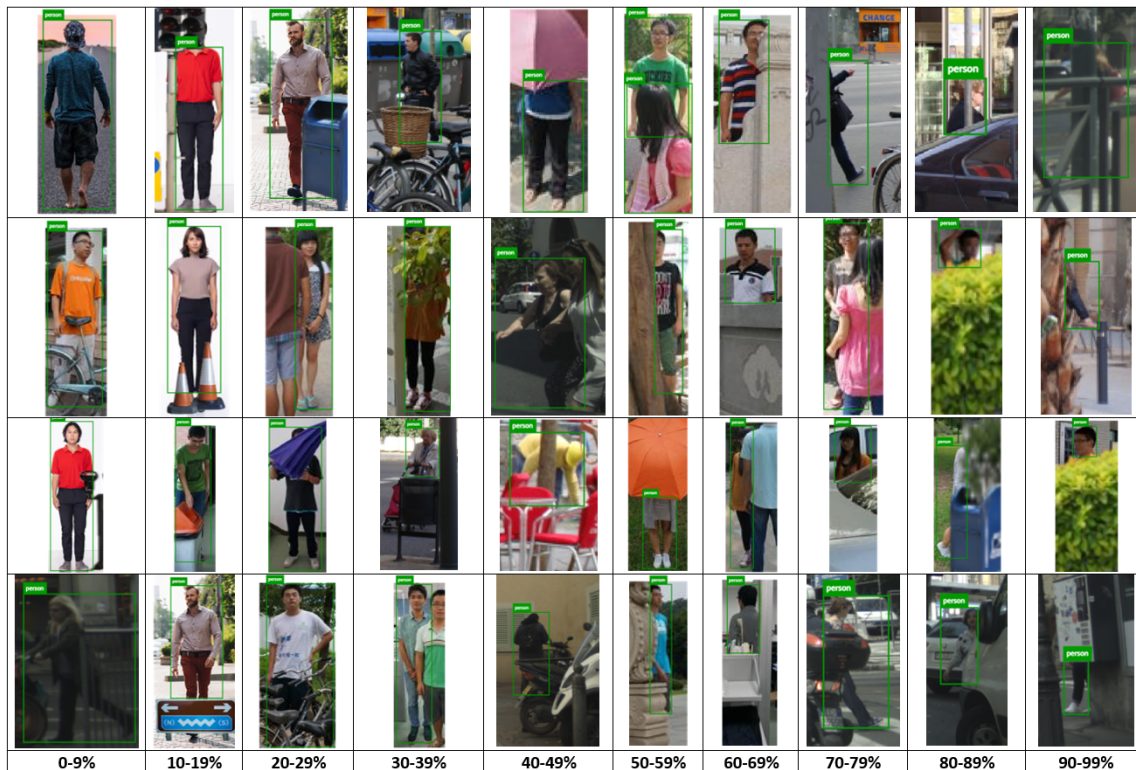


Figure 4.2: Dataset Sample. An example of dataset images for each level of occlusion. The custom dataset contains 820 pedestrian instances containing a wide range of pedestrian poses and occluding objects. All images are compiled from publicly available sources.

mation. Two-stage frameworks such as FasterRCNN [206], MaskRCNN [190] and R-FCN [207] apply two separate networks to perform classification. One network is used to propose regions of interest and a dedicated second network performs object detection [213]. One-stage frameworks such as RetinaNet [211], SSD [208] and SSDLite [209, 210] attempt to reduce computation and increase speed by performing object detection using a single feed forward convolutional network that does

Table 4.1: Overview of Pedestrian Detection Models.

Model	Classifier	Training Data	Weights Source	No. Parameters	Performance (mAP)
FasterRCNN [206]	ResNet-50 FPN	COCO	Voxel51 [205]	41.8 Million	0.398
MaskRCNN [190]	ResNet-50 FPN	COCO	Voxel51 [205]	44.4 Million	0.411
R-FCN [207]	ResNet-101	COCO	Voxel51 [205]	171.9 Million	0.411
SSD [208]	VGG16	COCO	Torchvision [204]	35.6 Million	0.412
SSDlite [209] [210]	MobileNetV3 Large	COCO	Torchvision [204]	3.4 Million	0.464
RetinaNet [211]	ResNet-50 FPN	COCO	Voxel51 [205]	34.0 Million	0.361
CenterNet [212]	Hourglass-104	COCO	Voxel51 [205]	189.3 Million	0.533

not interact with a region proposal module. RetinaNet also implements a novel method of “focal loss” which is used to reduce the imbalance between foreground and background classes during training with a view to increasing detection precision. CenterNet [212] takes an alternative approach based on keypoint estimation. Objects are represented as a single point at their bounding box center identified by a heat map generated using a fully convolutional network. Other object features such as object size, orientation and pose are then regressed directly from the image features at the center location. CenterNet has been shown to outperform a number of state of the art one-stage and two-stage algorithms in terms of a speed-accuracy trade off by maintaining an efficient network architecture. Further details of these experiments can be found in [212].

4.4.2 Experiments

Detection performance is analysed for the complete test dataset, and for each occlusion range from 0-9% to 90-99%, for pedestrian detection models to assess the impact of progressive levels of occlusion on the detectability of pedestrians. Analysis is carried out using Voxel51 [214] and the COCO style evaluation metric Mean Average Precision (mAP). Mean Average Precision is a popular and rigorous metric for object detection that calculates the Average Precision (AP) for a range of Intersection over Union (IoU) values from 0.5 to 0.95 with a step size of 0.05 and produces the mean value [189]. A summary of the results are shown in Figure 4.3, Figure 4.4 and Figure 4.5. All models are also characterised using the KITTI Vision Benchmark [7] in order to compare and demonstrate the advanced analysis capabilities provided by the proposed benchmark. Results on the KITTI Vision Benchmark are shown in Figure 4.8.

4.5 Results and Analysis

Results demonstrate that pedestrian detection performance (mAP) declines as the level of pedestrian occlusion increases, Figure 4.3. The number of false negative detections increase as occlusion level increases, Figure 4.4(c) and in general, the number of true positive detections begin to significantly decrease as occlusion level increases for pedestrians more than 50% occluded, Figure 4.4(a).

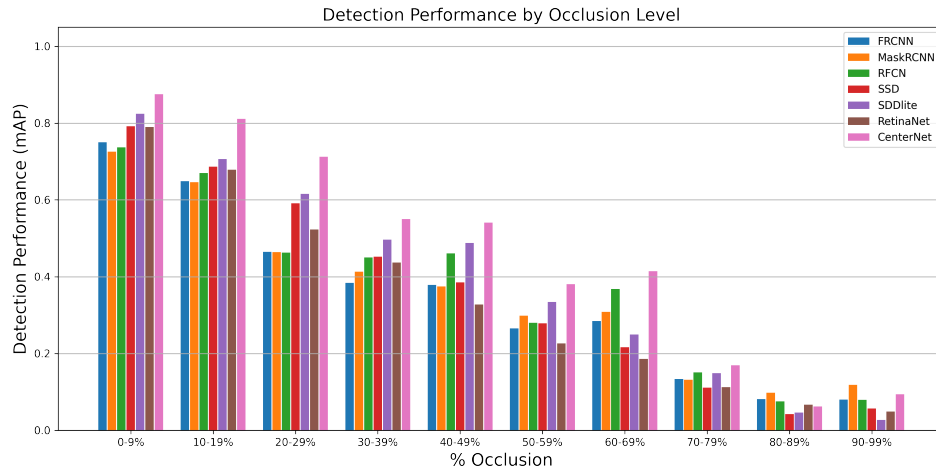
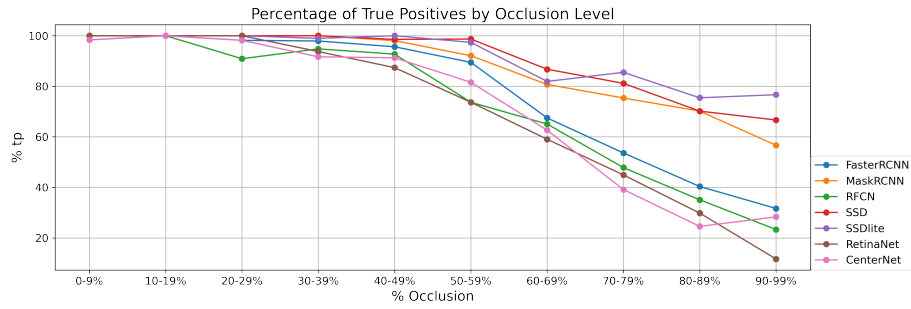
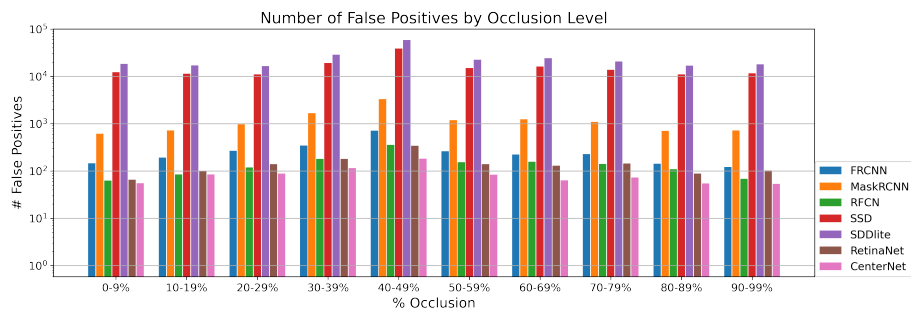


Figure 4.3: Detection Performance by Occlusion Level. Pedestrian detection performance of seven popular pedestrian detection models is displayed for images containing progressive levels of occlusion. Pedestrian detection performance (mAP) declines as the level of pedestrian occlusion is increased. CenterNet [212] is the highest performing detection model for pedestrians up to 80% occluded.

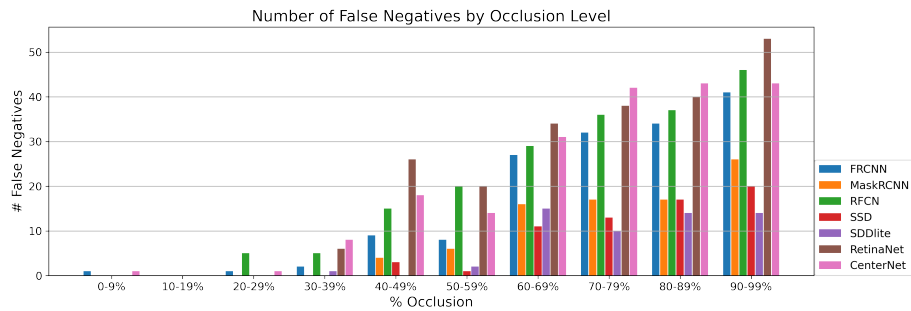
As shown in Figure 4.5, of the seven popular pedestrian detection models analysed, CenterNet [212] has the greatest overall detection performance for partially occluded pedestrians with an overall mAP of 0.533, followed by SSDLite [209, 210] with a total dataset mAP of 0.464. The strategy employed by CenterNet of first identifying the bounding box centre using a keypoint heatmap and then predicting object size and bounding box dimensions relative to the centre point has demonstrated the highest precision bounding boxes for both fully visible pedestrians and for pedestrians up to 80% occluded, Figure 4.3. MaskRCNN [190] has the greatest



(a)



(b)



(c)

Figure 4.4: True Positives, False Positives and False Negatives. (a) displays the percentage of true positive detections by occlusion level for seven popular pedestrian detection models. (b) displays the number of false positives per occlusion level for each model. Note the logarithmic scale on the Y-axis. (c) displays the number of false negatives by occlusion level.

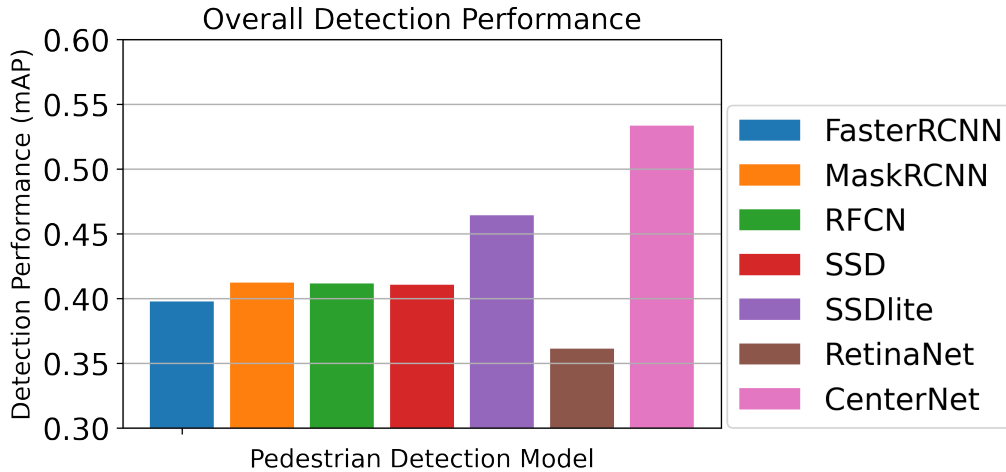


Figure 4.5: Overall Detection Performance. CenterNet has the greatest overall performance with a mAP of 0.533, followed by SSDlite (mAP = 0.464). RetinaNet has the lowest overall performance on the test dataset with a mAP of 0.361.

detection performance for pedestrians occluded more than 80%, Figure 4.3. RetinaNet [211] is the lowest performing overall on the test data with a mAP of 0.361. RetinaNet’s true positive detections begin to degrade in a linear fashion once pedestrians are more than 30% occluded and this model has the highest number of false negatives for pedestrians more than 30% occluded, Figure 4.4(a) and 4.4(c). Single Shot Detectors, SSD [208] and SSDLite [209, 210] have the highest number of true positive detections at high levels of occlusion, Figure 4.4(a), and maintain a very high level of true positive detections up to 60% occlusion, however their false positive rate is in the region of 100 times larger than popular two stage detectors such as FasterRCNN [206] and RFCN [207] and approximately 16 times larger than MaskRCNN

[190], Figure 4.4(b). Unlike false negatives, the number of false positives per image does not appear to be significantly impacted by the occlusion level as these are not typically related to the target pedestrian in an image. SSDlite [209,210] outperforms SSD [208] for almost all levels of occlusion despite having a higher number of false positive detections. MaskRCNN [190] has a higher percentage of true positives than Faster RCNN [206] for pedestrians over 40% occluded, however, it has around 4 times more false positive detections for the same data, Figure 4.4. Mask RCNN, RFCN and SSD all have similar overall performance on the test dataset, however, MaskRCNN and RFCN have a higher detection performance than SSD for pedestrians that are more than 60% occluded, Figure 4.3.

Figure 4.6 compares the output from a two stage detector, FasterRCNN, with a one stage detector, SSD, for an occluded pedestrian. Two stage detectors first generate key regions of interest before applying object detection, one stage detectors directly apply object detection to the entire image. Figure 4.6 demonstrates that for the same image, FasterRCNN produces 4 detection outputs (1 true positive with 88% confidence and 3 false positives), Figures 4.6(b) and 4.6(c), whereas SSD produces 84 detection outputs (1 true positive with 20% confidence and 83 false positives), Figures 4.6(d) and 4.6(e). These figures confirm that the characteristics and weaknesses of each detection model identified through robust performance characterisation, must be taken into account further downstream in a pedestrian detection system, as some model outputs may be less reliable than others for safety critical systems.

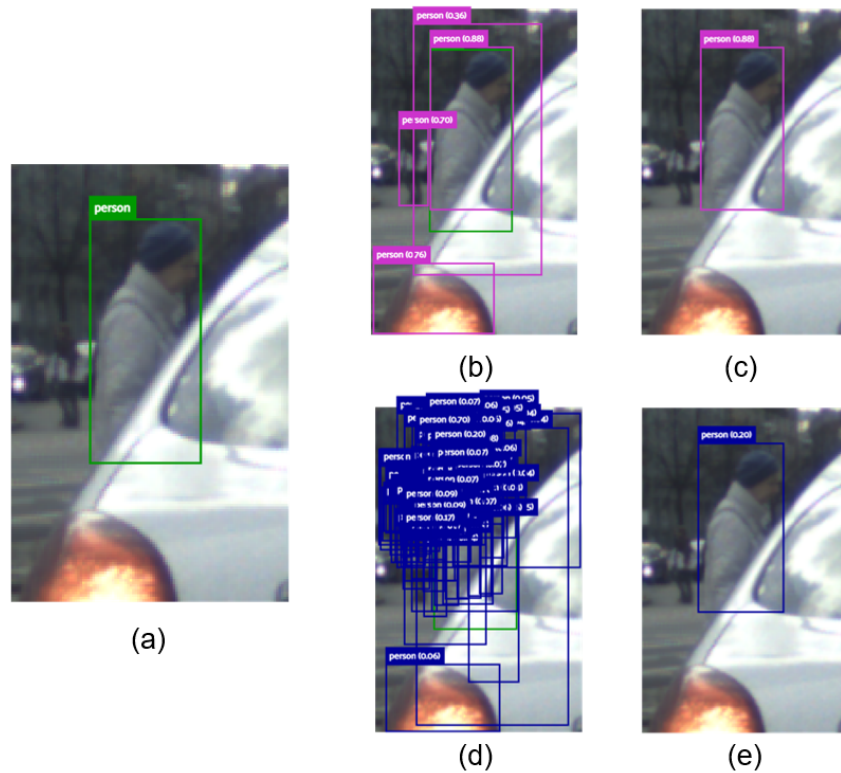


Figure 4.6: FasterRCNN vs. SSD. Detection performance is compared for a two stage network, FasterRCNN vs. a one stage network, SSD for an occluded pedestrian. The ground truth is shown in (a). FasterRCNN generates 4 proposals (b), 1 true positive detection with 88% confidence (c), and 3 false positives. SSD generates 84 detections (d), 1 true positive with 22% confidence (e), and 83 false positives.

4.5.1 Benchmark Comparison

Although a number of datasets contain occlusion labels to indicate the level of occlusion, current benchmarks are not designed for thorough characterisation of partially occluded pedestrian detection performance. Each benchmark varies greatly

in their definition of the occurrence and severity of occlusion and each benchmark uses different subjective methods of occlusion level annotation, Chapter 3, Table 2.3 [12]. In addition, many pedestrian instances are impacted by multiple additional inhibiting factors, making it difficult to determine if the contributing factor to non-detection is occlusion level alone. Algorithm performance can still be compared using the current state of the art, however users are unable to determine with any certainty if non-detection is the result of occlusion or one of many other inhibiting factors such as object scale, distance from camera, blur/focus, adverse weather and lighting variations. This also makes it very difficult to accurately compare algorithm performance across multiple benchmarks.

Taking the popular KITTI Vision Benchmark as an example. Images are annotated for three levels of occlusion: “Fully Visible”, “Partially Occluded”, “Difficult to See”. Images are captured using a wide angle lens and contain many contributing factors to non-detection in addition to occlusion as shown in Figure 4.7.

The dataset is split into three test subsets in order to characterise pedestrian detection models by occlusion label:

1. Images that *only* contain pedestrians tagged as “**Fully Visible**” (*1669 Instances in 1242 Images*)
2. Images that *only* contain pedestrians tagged as “**Partially Occluded**” (*236 Instances in 216 Images*)
3. Images that *only* contain pedestrians tagged as “**Difficult to See**” (*208 In-*

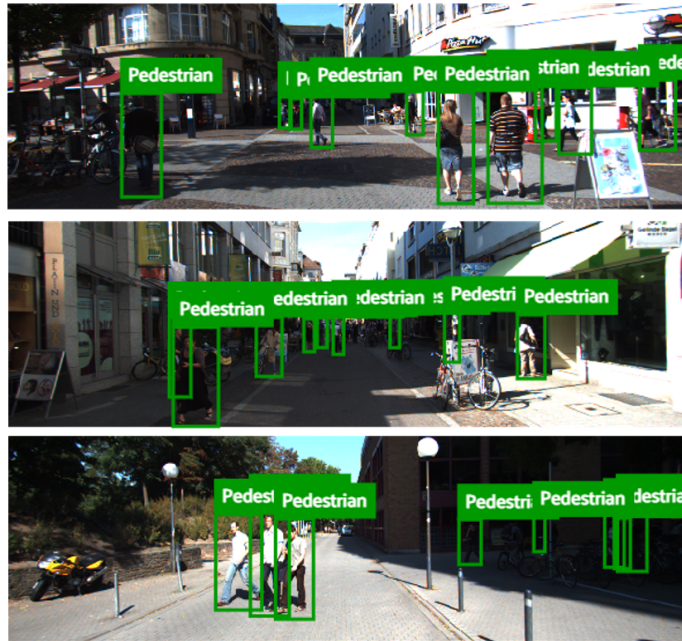


Figure 4.7: Example Images from the KITTI Vision Benchmark. Although pedestrians are tagged with occlusion level information, pedestrian instances are impacted by a range of inhibiting factors in addition to partial occlusion such as object scale and lighting variations. As a result it is difficult to ascertain the explicit impact of occlusion alone.

stances in 158 Images)

Note: sitting persons and persons on bicycles were included for test purposes in cases where they have a suitable occlusion label.

Pedestrian detection performance is then assessed on each of the three subsets as shown in Figure 4.8. Results demonstrate that performance declines for each data subset and MaskRCNN [190] has the greatest overall performance on the KITTI Vision Benchmark data. However, partial occlusion can not be concluded as the only

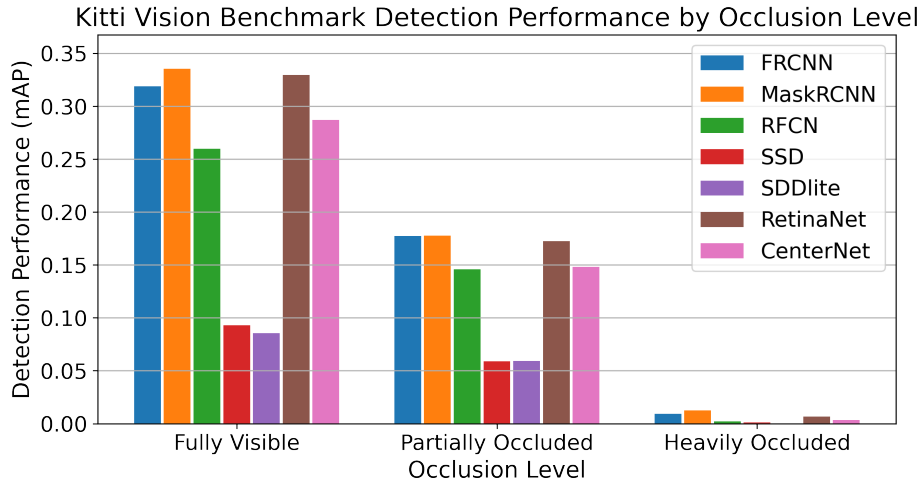


Figure 4.8: Detection Performance by Occlusion Level KITTI Vision Benchmark

contributing factor to non-detection as many pedestrian instances have a number of additional inhibiting factors.

In contrast to this, the proposed benchmark facilitates detailed, objective and repeatable characterisation of pedestrian detection performance specifically for partially occluded pedestrians across the complete range of occlusion levels from 0-99%, Figure 4.3.

4.5.2 Key Semantic Parts

Further analysis has been carried out to determine the impact that visibility of a pedestrian's head has on detection of occluded pedestrians. The dataset was split into two subsets: 1) Only images where the target pedestrian's head is visible and 2) Only images where the target pedestrian's head is occluded. Of the 820 pedestrian

instances, the target pedestrian’s head is visible in 582 instances and is occluded in 252 instances. Figure 4.9(a) displays the percentage of pedestrian instances with their head visible across each of the occlusion levels. Three pedestrian detection models, FasterRCNN, RetinaNet and SSD were then tested on both data subsets across the occlusion range.

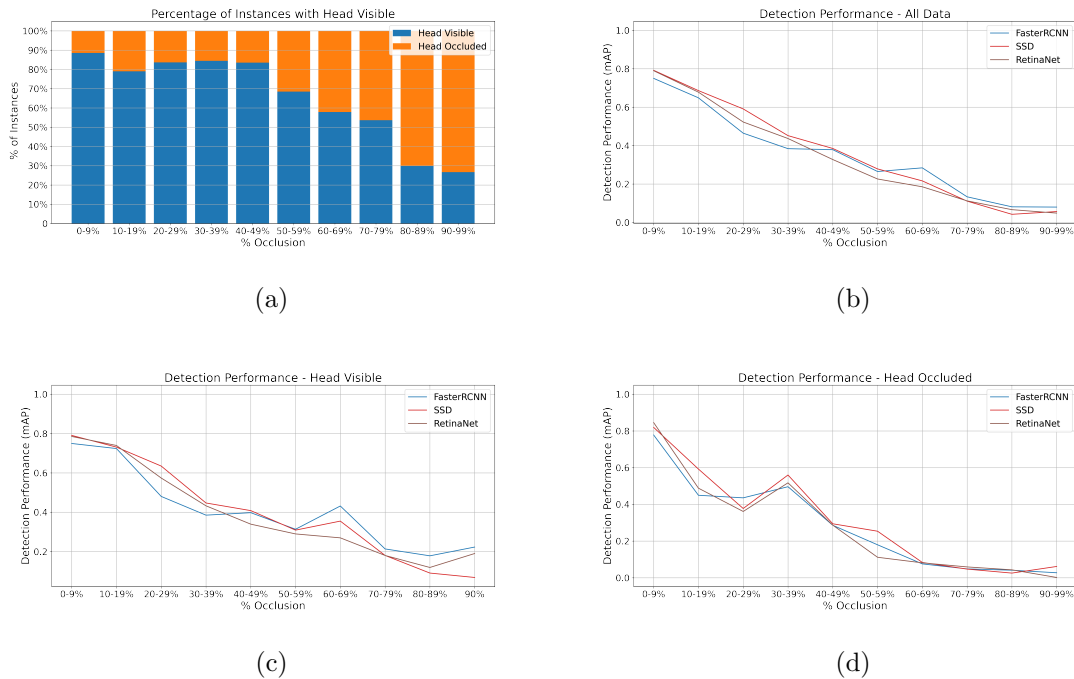


Figure 4.9: Analysis of data based on head visibility. (a) Dataset statistics based on head visibility. 820 total instances, 568 instances with head visible, 252 instances with head occluded. (b) Detection performance for FasterRCNN, SSD and RetinaNet for all data. (c) Detection performance for only images with head visible. (Note, no occlusion level of more than 90% possible with head visible.) (d) Detection performance for only images with head occluded.

Experiments demonstrate that, regardless of whether a pedestrian’s head is visi-

ble, a distinct declining profile in detection performance is observed as pedestrian occlusion level increases, Figures 4.9(b), 4.9(c) and 4.9(d). Figure 4.9 demonstrates the fine-grained analysis capability provided by the occlusion level classification method proposed in Chapter 3. Results indicate that the detection models under test contain a bias towards head visibility with a 0.27 reduction of mAP for FasterRCNN and 0.25 reduction of mAP for RetinaNet in the 10-19% occluded range for instances where the head is occluded compared to the same range where the head is visible. Further analysis of algorithm performance for the occlusion of specific semantic parts can provide insight into the impact of self occlusion on detection performance on a model-by-model basis.

4.6 Conclusion

Detection of partially occluded pedestrians remains a persistent challenge for driver assistance systems and autonomous vehicles. Current methods of characterising detection performance for partially occluded pedestrians have been broad, subjective, and inconsistent in their definition of the level of occlusion. This chapter presents a novel test benchmark for the detailed, objective analysis of pedestrian detection models for partially occluded pedestrians. Detection performance is characterised for seven popular pedestrian detection models across a range of occlusion levels from 0-99%. The proposed benchmark focuses specifically on the complex issue of partial occlusion and facilitates more objective, repeatable and fine grained

analysis than the current state of the art. Results demonstrate that pedestrian detection performance declines as the occlusion level increases and the visibility of a pedestrian is reduced. An increase in the number of false negative detections is observed as occlusion level increases and the percentage of true positive detections significantly degrade for pedestrians who are more than 50% occluded. Further analysis demonstrates that not all pedestrian detection models should be treated equally within an object detection system. The speed vs. accuracy trade-off, encouraged by the time sensitive detection requirements of autonomous vehicles, can result in high levels of false positive detections and lower detection confidence at progressive levels of pedestrian occlusion, particularly when using single stage detection models. Thorough objective characterisation of pedestrian detection models at the design stage will improve the performance of object detection systems by calibrating the priority of detections in scenarios where known weaknesses can occur. System improvements may be gained through the use of an occlusion-aware step in the object detection pipeline to inform the priority of camera-based detections in sensor fusion networks for SAE level 4 and level 5 autonomous vehicles. In this manner, any reduction in performance at high occlusion levels can be mitigated in the design of the overall system to increase the safety of vulnerable road users and improve the efficiency of path planning based on environment detection. Widespread use of the proposed benchmark can result in more objective, consistent and detailed analysis of pedestrian detection models for partially occluded pedestrians.

Chapter 5 further progresses this research theme by synthesising and applying the learning outcomes and research contributions of Chapter 2, Chapter 3 and Chapter 4 to develop an objective benchmark for e-scooter rider detection and to inform the development of a novel, occlusion-aware method of e-scooter rider detection.

Chapter 5

E-Scooter Rider Detection and Classification in Dense Urban Environments

5.1 Summary

Accurate detection and classification of vulnerable road users is a safety critical requirement for the deployment of autonomous vehicles in heterogeneous traffic. Although similar in physical appearance to pedestrians, e-scooter riders follow distinctly different characteristics of movement and can reach speeds of up to 45kmph. The challenge of detecting e-scooter riders increases in urban environments where the frequency of partial occlusion is increased. This can lead to non-detection or

mis-classification of e-scooter riders as pedestrians, providing inaccurate information for accident mitigation and path planning in autonomous vehicle applications. This chapter introduces a novel benchmark for partially occluded e-scooter rider detection to facilitate the objective characterisation of detection models. A novel, occlusion-aware method of e-scooter rider detection is presented that achieves a 15.93% improvement in detection performance over the current state of the art.

5.2 Introduction

Micro-mobility solutions such as e-scooters have seen a rapid rise in popularity in recent years as many cities seek modern solutions to ease traffic, emissions and parking difficulties in built up areas [215]. The intuitive operation of e-scooters, and the growing number of service providers offering short term rentals, have prompted market predictions that shared e-scooter usage may ultimately capture 8-15% of all trips shorter than 5 miles, worldwide [215]. The proliferation of e-scooters adds an additional level of complexity to the detection and classification of vulnerable road users. Although very similar in physical appearance, e-scooter riders and pedestrians behave very differently in the automotive environment. E-scooter riders can reach speeds of up to 45 kilometres per hour [110–112] and follow distinctly different movement characteristics than pedestrians. The challenge of accurately detecting and classifying e-scooter riders is more complex in urban environments as the frequency and severity of partial occlusion is increased. This can lead to the non-detection or

mis-classification of e-scooter riders as pedestrians or other road users, providing inaccurate information for accident mitigation and path planning. In addition, recent research indicates that e-scooter usage is currently one of the most dangerous forms of transportation with 115 injuries per million trips [216], substantially higher than motorcycles (104 injuries per million trips), bicycles (15 injuries per million trips) and walking (2 injuries per million trips) [217].

Leading pedestrian and cyclist detection systems claim a detection performance of approximately 65%-75% of partially and heavily occluded instances respectively using current benchmarks, Chapter 2 [10, 11, 184–186]. However, very few research articles exist on the safety critical challenge of e-scooter rider detection to date and to the best of the authors knowledge, no known research has been carried out on the detection and classification of e-scooter riders under partial occlusion. A knowledge gap exists for an objective benchmark for e-scooter rider detection performance in urban areas where the frequency and severity of partial occlusion is increased.

This chapter presents a novel, objective benchmark for partially occluded e-scooter riders to facilitate the characterisation of vulnerable road user detection and classification models. A novel, occlusion-aware method of e-scooter rider detection is presented and objective performance characterisation is carried out for a range of popular classifiers for the complete spectrum of occlusion levels from 0-99%. The contributions of this research are: 1. A novel, objective, test benchmark for partially occluded e-scooter rider detection and classification is presented. 2. A novel,

occlusion-aware method of e-scooter rider detection is described which provides a 15.93% improvement on the current state of the art e-scooter rider detection network as described in [168]. 3. Objective characterisation of e-scooter rider classification is carried out for a number of popular, publicly available classifiers.

5.3 Related Work

Limited research exists on e-scooter rider detection to date. Apurv *et al* [168] present a baseline algorithm for e-scooter rider detection. Candidate selection is carried out using YoloV3 with pre-trained weights [218] on the COCO dataset [189]. The bounding box dimensions for each person instance are then enlarged on three sides (left, bottom and right) using the formula outlined in Equation 5.1, to incorporate the surrounding area where an e-scooter is normally located in instances where the detected person is an e-scooter rider.

$$(x, y, w, h) \Rightarrow ((x - w), y, 3w, (h + h/4)) \quad (5.1)$$

Where "x" and "y" represent the x-axis and y-axis coordinates of the top left corner of the bounding box; "w" = bounding box width and "h" = bounding box height.

The extended bounding box regions are then fed into a MobileNetV2 classifier [210], trained on the "IUPUI CSRC E-Scooter Rider Detection Benchmark Dataset" [168]. The IUPUI E-Scooter Rider Dataset contains 21,454 images for binary classi-

fication including 10,749 images containing e-scooter riders and 10,705 images which do not contain an e-scooter rider. The authors claim a validation accuracy of more than 0.9, however very few instances of occluded e-scooter riders are included in the validation data and no reference is made as to the ability of the network to generalise to new data.

Nguyen *et al* [169] also utilise YoloV3 to implement an e-scooter rider detection system, however this approach focuses on detecting an e-scooter and its rider as two separate classes. The methodology separates the image into a grid and relates adjacent bounding boxes of the target classes in order to identify e-scooter riders. The network is trained using 140 training images and 60 validation images obtained through searches on Baidu and Google Images using the keyword “rider and scooter”. Transfer learning is then used to fine tune the YoloV3 model to the target classes. The authors expand this research by exploiting the detection of two separate classes to identify cases where the detected person is horizontal to the e-scooter, indicating a potential fall. The authors also claim a validation accuracy of over 0.9, however only 60 validation images are used, no instances of occluded e-scooter riders are included and no reference is made to more thorough evaluation indicating the networks’ ability to generalise to new data.

Researchers at the Digital Transformation Hub at California Polytechnic State University collaborated with the City of Santa Monica in 2018 to implement a machine learning based e-scooter detection and counting system, in order to help mon-

itor and enforce the prevention of e-scooter riding on sidewalks [171]. E-scooter rider detection was achieved through transfer learning of a pre-trained RetinaNet object detection algorithm using an in-house custom dataset. A parallel Resnet50 semantic segmentation branch was also used to differentiate between the sidewalk and the road surface. Overlapping e-scooter rider and sidewalk detections indicate an infringement and the instance is counted and tracked for enforcement purposes [170].

Many popular pedestrian detection benchmarks provide occlusion level annotation to determine the relative detection performance for partially occluded pedestrians [6–9, 38, 102, 105, 113, 114, 117, 118, 194]. Although less represented, there are also a significant number of cyclist detection benchmarks with occlusion specific annotation [7, 9, 102, 194], however, no known e-scooter detection benchmark with occlusion labels exists to date.

Chapter 4 [13] presents an objective benchmark for partially occluded pedestrian detection, containing 820 pedestrian instances under progressive levels of occlusion from 0-99%. Images are annotated using the objective method of occlusion level annotation described in Chapter 3 [12]. Keypoint detection is used to identify semantic body parts and findings are cross-referenced with a visibility score and the pedestrian mask in order to confirm the presence or occlusion of each semantic part. A novel method of 2D body surface area estimation based on the “Wallace rule of Nines” [11, 125] is then used to calculate the total occlusion level of each pedestrian

instance. Inspired by the research described in Chapter 3 [11, 12] and Chapter 4 [13], this research uses a novel objective benchmark for e-scooter rider detection. In contrast to prior works [168, 169] the proposed benchmark itemises algorithm performance for partially occluded e-scooter riders for the complete range of occlusion levels from 0%-99%. In addition, a novel, occlusion-aware e-scooter rider detection network is described to improve upon the current state of the art.

5.4 Methodology

A novel e-scooter rider test dataset, containing 1,130 images including 543 e-scooter rider instances and 587 other vulnerable road user instances, has been created in order to characterise e-scooter rider detection and classification across a range of occlusion levels from 0 to 99% occluded. A diverse mix of images are used to ensure that a wide variety of e-scooter riders, orientations, backgrounds, and occluding objects are represented.

The dataset is compiled from publicly available, web crawled sources. Occluding objects are superimposed on to e-scooter riders with progressive levels of occlusion. This dataset is then complemented by 587 instances of pedestrians and cyclists across an identical range of occlusion levels. Non e-scooter rider images are sourced from the occluded pedestrian detection dataset presented in Chapter 4. All images are annotated using the objective occlusion level classification method described in Chapter 3 [12]. Complex cases at very high occlusion rates were manually verified using the

method of 2D body surface area estimation presented in Chapter 3. Dataset statistics by occlusion level and a sample of the test dataset can be seen in Figure 5.1 and Figure 5.2 respectively.

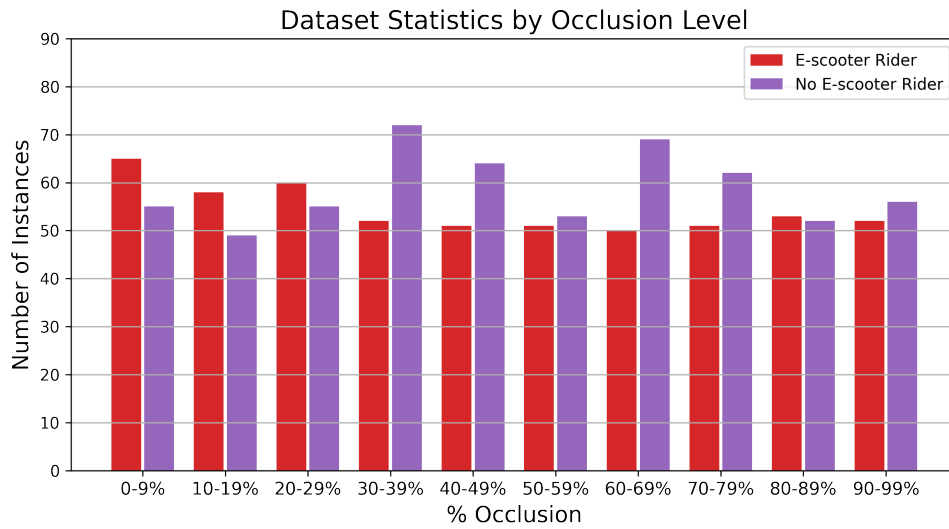


Figure 5.1: Test Dataset Statistics. The number of target instances per occlusion level. The custom dataset contains 1,130 images under progressive levels of occlusion from 0-99%.



Figure 5.2: Test Dataset Sample. An example of dataset images for each level of occlusion. The custom dataset consists of 1,130 images, including 543 e-scooter rider images and 587 non e-scooter rider images.

5.4.1 E-Scooter Rider Classification

Classifier performance is evaluated using the total test dataset for the current state of the art, as outlined in [168], and for five popular, publicly available classifiers in order to compare performance across the complete range of occlusion levels. Each classifier, AlexNet [219], SqueezeNet1.0 [220], VGG16 with Batch Normalisation (VGG16_bn) [221], ResNet34 and ResNet101 [99] is trained on the IUPUI E-Scooter Rider Dataset [168] using Pytorch [222] and Fast AI [223]. The detection and classification pipeline proposed in [168] is used to maintain consistency and provide baseline results for comparison. Analysis is carried out using Voxel51 [214] and COCO style evaluation metrics. Accuracy is calculated using the formula highlighted in Equation 5.2, where TP = Number of true positives, TN = Number of true negatives, FP = Number of false positives and FN = Number of false negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.2)$$

A comparison of classifier performance using the methodology outlined in [168] can be seen in Figure 5.5.

5.4.2 Occlusion-Aware E-Scooter Rider Detection

A novel, occlusion-aware method of e-scooter rider detection is proposed to increase the performance of e-scooter detection in heterogeneous traffic.

1. Potential e-scooter rider instances are detected using a CenterNet-Hourglass104 [212] based, COCO trained person detector. CenterNet-Hg104 has been selected for region of interest generation due to the model’s high detection performance for occluded pedestrians in the research and experiments carried out in Chapter 4.
2. The aspect ratio of each bounding box is then analysed to determine if the detected instance is likely to be occluded. Detected bounding boxes of all potential candidates are then expanded on 3 sides as outlined in Figure 5.4. The extent to which the bounding boxes are expanded is based on the aspect ratio of the initial detection. If the bounding box height is less than 2.5 times the bounding box width, the height of the bounding box is increased by a higher magnitude to incorporate the pixel area where an e-scooter would be located in normal operation.
3. Modified bounding boxes are then processed by a custom trained ResNet101 classifier to classify instances of e-scooter riders. The classifier is trained using the e-scooter rider dataset presented in [168]. The dataset contains 21,454 training images for binary classification, consisting of 10,749 “e-scooter rider” images and 10,705 “non e-scooter rider” images.

An example of the efficacy of this method for partially occluded e-scooter users, compared to the current state of the art, can be seen in Figure 5.3. A flowchart of

the proposed occlusion-aware detection pipeline can be seen in Figure 5.4.

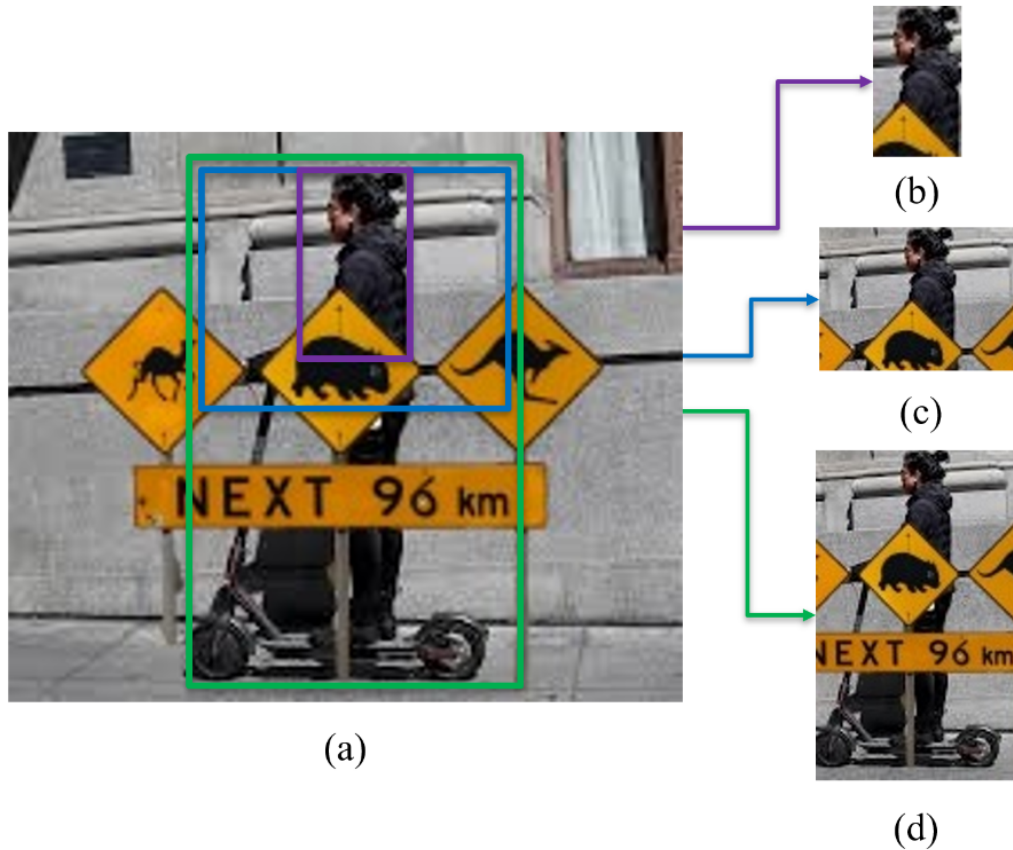


Figure 5.3: Candidate Selection Output Comparison. Example of the efficacy of the proposed candidate selection output for partially occluded e-scooter riders. The input image is displayed in (a). The cropped bounding box area from the initial detection algorithm is shown in (b). The cropped bounding box from the current state of the art as presented in [168], is shown in (c). The cropped bounding box area for the proposed novel, occlusion-aware e-scooter rider detection method is displayed in (d). The proposed method more comprehensively incorporates the e-scooter for partially occluded instances than the prior state of the art.

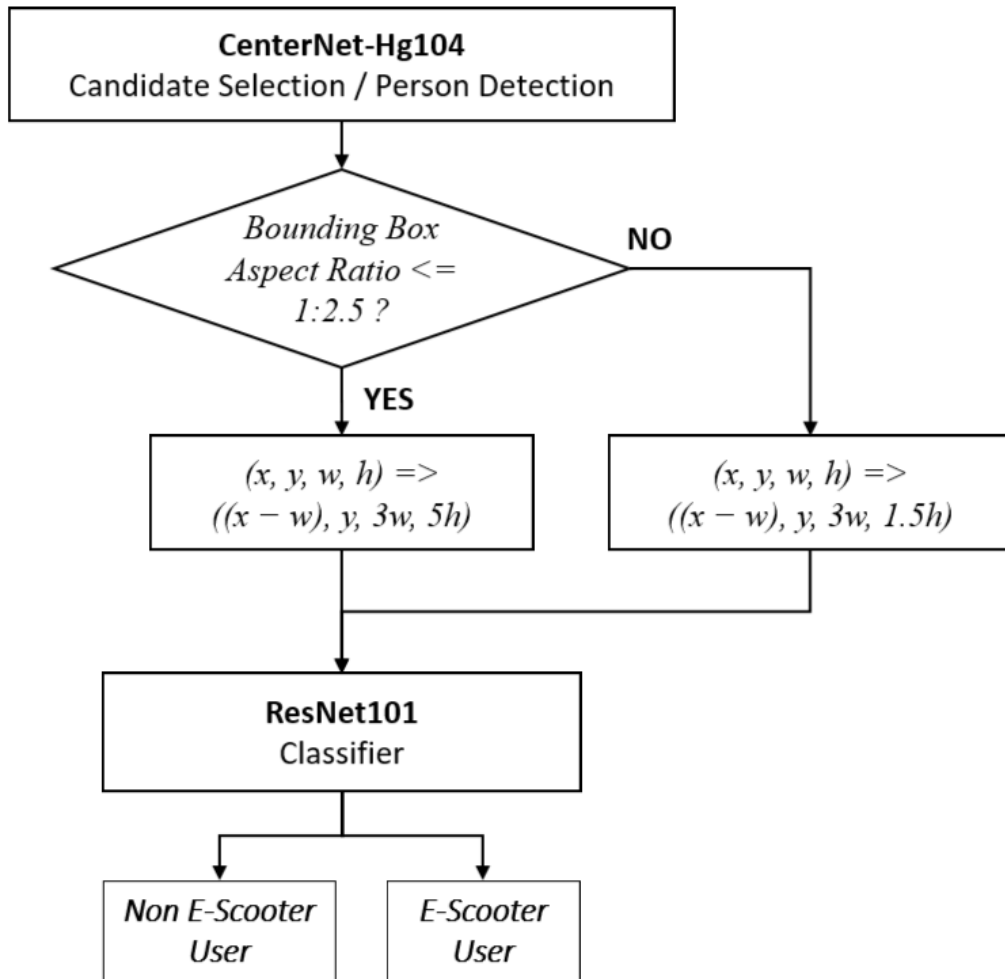


Figure 5.4: Occlusion-Aware E-Scooter Detection Flowchart.

5.4.3 Performance Characterisation

Detection and classification performance is characterised for e-scooter riders and other vulnerable road users for the complete test dataset and for each level of occlusion from 0-9% to 90-99%. The detection method proposed in Section 5.4.2 is compared to the current state of the art [168], and to four additional classifier configurations based on the proposed pipeline. All classifiers, AlexNet [219], SqueezeNet1.0 [220], VGG16 with Batch Normalisation (VGG16_bn) [221], ResNet34 and ResNet101 [99], are trained using the e-scooter rider dataset presented in [168]. The overall detection performance of each network is shown in Figure 5.5 and Figure 5.6. Detailed characterisation of the detection performance for each level of occlusion is presented in Figure 5.7 and Figure 5.8.

5.5 Results and Analysis

Figure 5.5 compares the performance of five popular classification networks based on the methodology outlined by the current state of the art [168]. Results demonstrate that for a mixed occlusion dataset, ResNet101 and ResNet34 [99] achieve a 2.1% improvement over the MobileNetV2 classifier [210] used by Apurv *et al* [168], using the same training data, backbone detection network, and classification pipeline.

A novel occlusion-aware method of e-scooter rider detection is described in Section 5.4.2. Experiments show that the proposed methodology is more proficient at

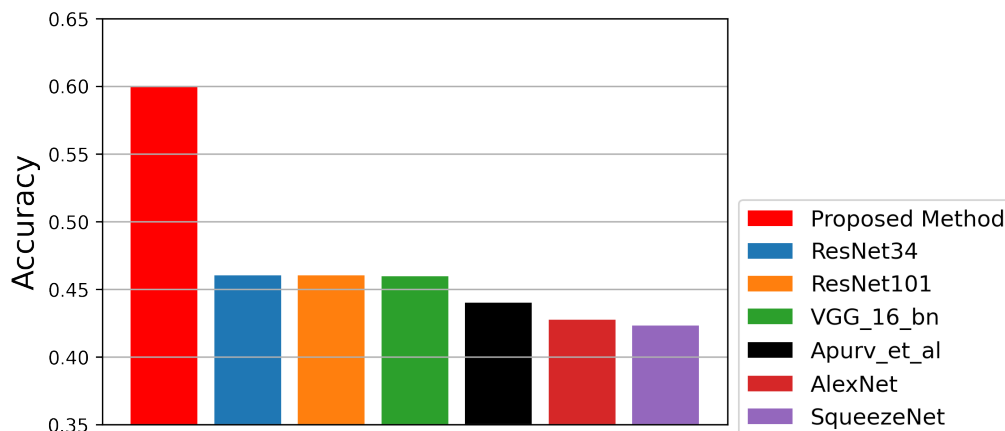


Figure 5.5: Detection and Classification Performance. Detection and Classification performance for the total test dataset using the methodology outlined by the current state of the art [168]. ResNet101 and ResNet34 are the highest performing classifiers, each with a test accuracy of 0.460. The baseline method proposed by Apurv *et al* [168] has a test accuracy of 0.439. Results demonstrate that the proposed detection network achieves an accuracy improvement of 15.93% over the current state of the art.

detecting partially occluded e-scooter riders with an overall accuracy of 0.599, a 15.93% improvement over the current state of the art [168], Figure 5.5. Detailed results of the detection accuracy and the percentage of true positives for each occlusion level are shown in Figure 5.7a and Figure 5.7b respectively. The number of false negatives by occlusion level is shown in Figure 5.8. Characterisation results show that for each level of occlusion, the proposed method provides a superior detection accuracy, a higher percentage of true positives and a lower percentage of false negatives than the current state of the art [168], Figure 5.7 and Figure 5.8. Results also demonstrate that, in general, e-scooter detection accuracy, and the percentage of true

positives decline as occlusion level increases, and the percentage of false negatives increase with occlusion level. This reflects the findings of Chapter 4 [13] and presents a significant challenge when detecting and classifying e-scooter riders in dense urban environments where the frequency and severity of partial occlusion is increased.

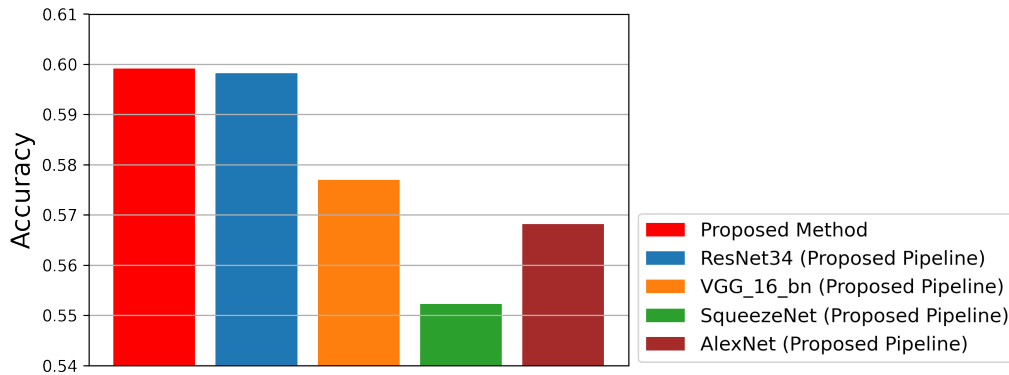
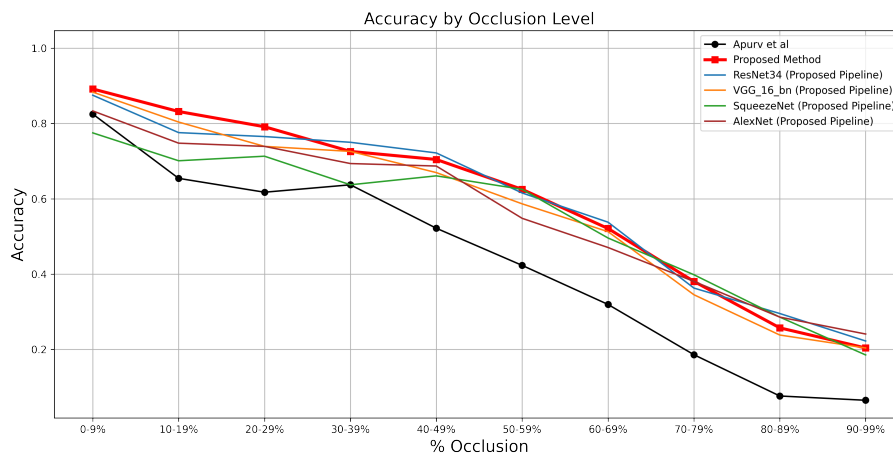
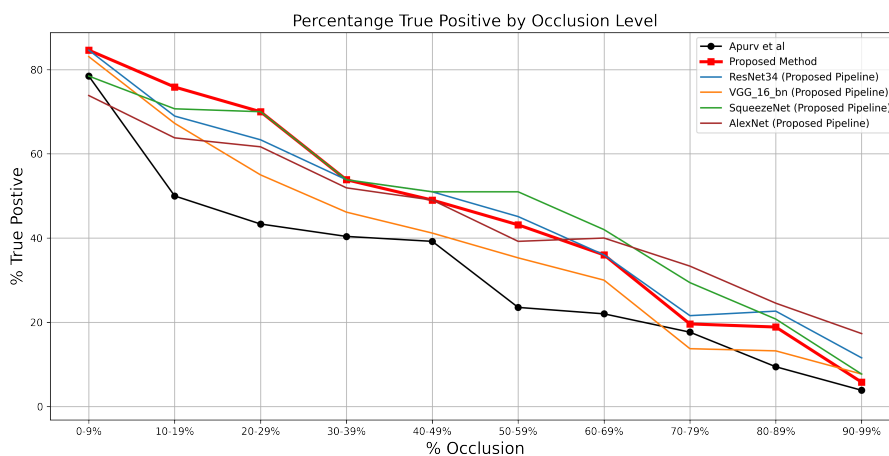


Figure 5.6: Classifier Comparison using the Proposed Occlusion-Aware Pipeline. The proposed e-scooter rider detection network is compared to four alternative classifier configurations using the proposed pipeline. The ResNet101 classifier specified by the proposed method achieves the highest classification performance with an accuracy of 0.599.

Thorough characterisation of a detection algorithm at the system development stage can help identify the suitability of specific classification models for particular scenarios and applications. For example, further analysis demonstrates that Vgg16_bn [221] has a below average true positive rate, and an above average false negative rate, Figure 5.7b and Figure 5.8 respectively. However, VGG16_bn also maintains a relatively low number of false positive detections across the range of occlusion levels, resulting in the third most accurate classification performance overall, Figure



(a)



(b)

Figure 5.7: Detection Performance by Occlusion Level. The detection accuracy by occlusion level, (a), and the percentage of true positives per occlusion level, (b), is shown for the current state of the art, the proposed method and for a number of alternative classifier configurations using the proposed pipeline.

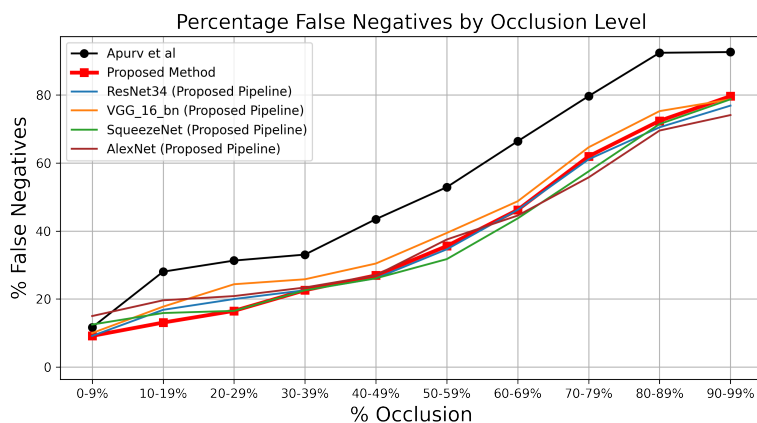


Figure 5.8: False Negatives by Occlusion Level. The percentage of false negatives per occlusion level is shown for the current state of the art, the proposed method and for a number of alternative classifier configurations using the proposed pipeline. The proposed method (red) consistently achieves a lower percentage of false negatives than the current state of the art (black) [168].

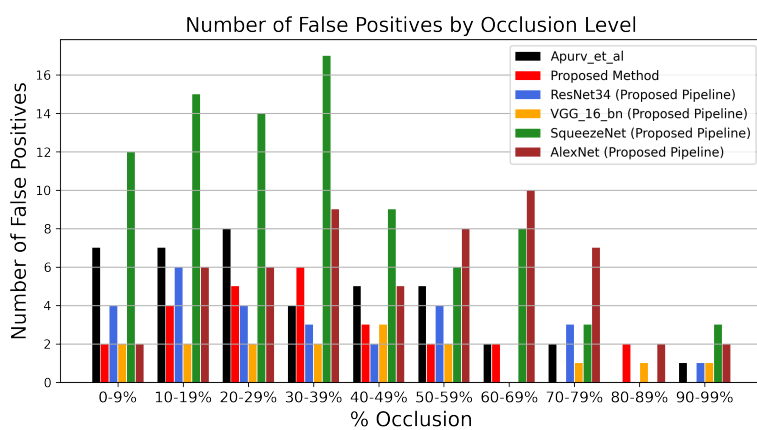


Figure 5.9: Number of False Positives by Occlusion Level. SqueezeNet1.0 [220] detects the highest number of false positives across the range of occlusion levels (87 total false positives), followed by AlexNet [219] (57 total false positives).

5.6. This provides insight into the selectivity of the network and the relatively lower confidence assigned to borderline detection instances. SqueezeNet1.0 [220] has a higher number of true positive detections for e-scooter riders who are between 40% and 60% occluded. AlexNet [219] achieves a higher percentage of true positives for instances that are more than 60% occluded, Figure 5.7b. However, both networks incur a significantly higher false positive rate across the range of occlusion levels, Figure 5.9. This is an important distinction as the mis-classification of e-scooter riders as pedestrians or vice versa, can result in dangerous scenarios in autonomous vehicle applications, such as the inappropriate application of emergency braking, potentially resulting in collisions from behind, erratic swerving or the unnecessary triggering of other accident mitigation routines.

5.6 Conclusion

The non-detection, or mis-classification of e-scooter users as pedestrians or other road users will have a significant impact on the accident mitigation capabilities and the safe navigation of smart, connected and autonomous vehicles. This research presents an objective test benchmark for the characterisation of detection models for partially occluded e-scooter riders. The novel, occlusion-aware e-scooter rider detection method described in this article achieves a 15.93% improvement in detection accuracy over the current state of the art as presented in [168], Figure 5.5. Detailed characterisation of the proposed method, and the current state of the art, is provided

for the complete range of occlusion levels from 0-99% occluded, Figure 5.7.

Chapter 6 concludes the thesis by providing a summary of the work carried out throughout the PhD project, an overview of the key conclusions from the research and discussing the future research opportunities available through further development of the concepts described in the thesis.

Chapter 6

Conclusions and Future Work

6.1 Project Summary and Conclusions

Accurate and robust detection of vulnerable road users is a safety critical requirement for the deployment of autonomous vehicles in heterogeneous traffic. One of the most complex outstanding challenges is that of partial occlusion where a target object is only partially available to the sensor due to obstruction by another foreground object. The frequency and variety of occlusion in the automotive environment is large and diverse as pedestrians, e-scooter riders and cyclists navigate between vehicles, buildings, infrastructure and other road users. This thesis focuses on the detection of partially occluded pedestrians and e-scooter riders in the automotive environment.

Chapter 2 provides a thorough literature review of occlusion handling for vehicle detection, vulnerable road user detection and object detection in the automotive

environment. The literature review identifies a number of shortfalls and knowledge gaps in the current state of the art for partially occluded VRU detection in the automotive environment, including:

- A considerable amount of work has yet to be carried out on pedestrian and cyclist detection which is still only in the region of 65% to 75% detectable under partial occlusion using current benchmarks.
- No definitive metric or annotation methodology exists for the occurrence and severity of partial occlusion. As a result there is a large amount of inconsistency between current benchmarks.
- A significant knowledge gap exists for the detection of e-mobility users such as e-scooter riders under partial occlusion.

Chapter 3 begins to address these knowledge gaps through the development of an objective metric and methodology for pedestrian occlusion level classification for ground truth annotation. The proposed method uses keypoint detection and mask segmentation to identify and determine the visibility of the semantic parts of partially occluded pedestrians and calculates the percentage occluded body surface area using a novel, effective method for 2D body surface area estimation. The proposed method removes the subjectivity of the human annotator used by the current state of the art, in turn increasing the robustness and repeatability of pedestrian occlusion level classification. Qualitative and quantitative validation demonstrates the

effectiveness of the proposed method for all forms of occlusion including challenging edge cases such as self-occlusion and inter-occluding pedestrians. Experiments show a significant improvement over the current state of the art when plotted against the pixel-wise pedestrian occlusion level. Results demonstrate that the proposed method more closely reflects the pixel-wise occlusion level with a Root Mean Squared Error (RMSE) of 4.68 and Variance (VAR) of 21.88, compared to the current state of the art (RMSE = 18.09, VAR = 249.21).

Chapter 4 further develops this research through the production of an objective test dataset for benchmarking pedestrian detection performance for the complete spectrum of occlusion levels from 0-99%. The proposed benchmark provides fine-grained occlusion level characterisation for ten objectively defined occlusion ranges in contrast to the 2-3 broad categories used by the current state of the art. Performance characterisation is carried out for seven popular pedestrian detection routines to determine the impact of progressive levels of occlusion on pedestrian detectability. Additional experiments are conducted to determine the saliency of head visibility on detectability and to compare performance with a current state of the art pedestrian detection benchmark. Results demonstrate that the proposed benchmark provides more objective, detailed analysis capabilities for detection networks for partially occluded pedestrians than the current state of the art.

Chapter 5 synthesizes and applies the knowledge gained throughout the PhD research to improve upon the current state of the art for a modern, increasingly

popular category of vulnerable road user, the e-scooter rider. A novel objective benchmark for e-scooter rider detection is created using the principles gained in Chapter 3 and Chapter 4. Performance characterisation is carried out for the leading state of the art e-scooter detection algorithm. A novel, occlusion-aware method of e-scooter rider detection is described that achieves a 15.93% improvement in detection performance over the current state of the art.

The thesis draws a number of conclusions from the research on partially occluded vulnerable road user detection, including:

1. The detection of partially occluded pedestrians remains a persistent and underdeveloped challenge for driver assistance systems and autonomous vehicles. Partial occlusion is a frequent occurrence in the automotive environment and the occlusion ratio is demonstrated to have a direct impact on the detectability of pedestrians. Results of experiments described in Chapter 4 demonstrate that pedestrian detection performance declines as the occlusion level increases and the visibility of a pedestrian is reduced. An increase in the number of false negative detections is observed as the occlusion level increases and the percentage of true positive detections significantly decreases for pedestrians who are more than 50% occluded.
2. Current methods of characterising detection performance for partially occluded pedestrians have been broad, subjective, and inconsistent in their definition of the level of occlusion. The objective metric and method of occlusion level an-

notation proposed in this thesis can provide detailed, objective characterisation of detection performance for the complete range of occlusion levels from 0-99%.

3. Current pedestrian detection benchmarks do not differentiate between partially occluded e-scooter riders and pedestrians. This can present significant challenges in autonomous vehicle applications as the detection output is used to inform path planning and accident mitigation. Although partially occluded e-scooter riders appear similar to pedestrians from a perception point of view, e-scooters can reach speeds of up to 45kmph and demonstrate very different characteristics of movement in the automotive environment. Accurate classification, in addition to detection, is of increased importance in cases where road user classes share similar visual characteristics. Modern vulnerable road user detection benchmarks must have the flexibility to evolve and more readily incorporate new mobility solutions.
4. Considerable progress has been made in recent years using deep learning based detection networks due to the contrasting characteristics of the traditional road user classes: vehicle, pedestrian, motorcycle etc. However, the recent popularity of e-scooters presents a distinct challenge to partially occluded VRU detection systems based on deep learning alone. The convergence of deep learning and traditional computer vision based processing can provide significant gains in cases where detection classes share a large percentage of visual and pose characteristics such as pedestrians and e-scooter riders.

6.2 Primary Contributions

The primary contributions of this thesis can be summarised as follows:

- A comprehensive literature review on the theme of occluded object detection in the automotive environment as published in *S. Gilroy, E. Jones, and M. Glavin, “Overcoming occlusion in the automotive environment-a review”, IEEE Transactions on Intelligent Transportation Systems, 2019. [10]*.
- A novel, objective metric and methodology for pedestrian occlusion level classification for ground truth annotation as published in *S. Gilroy, M. Glavin, E. Jones, and D. Mullins, “Pedestrian occlusion level classification using key-point detection and 2d body surface area estimation”, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3833–3839 [11]* and *S. Gilroy, M. Glavin, E. Jones, and D. Mullins, “An objective method for pedestrian occlusion level classification”, Pattern Recognition Letters, 2022. [12]*.
- A novel, objective, test benchmark for partially occluded pedestrian detection as published in *S. Gilroy, D. Mullins, A. Parsi, E. Jones, and M. Glavin, “Replacing the human driver: An objective benchmark for occluded pedestrian detection” Biomimetic Intelligence and Robotics, 2023. [13]*.
- A novel, objective, test benchmark for partially occluded e-scooter rider detection and classification as published in *S. Gilroy, D. Mullins, E. Jones, A.*

Parsi, and M. Glavin, "E-scooter rider detection and classification in dense urban environments", Results in Engineering, Vol.16, 2022. [14].

- A novel, occlusion-aware method of e-scooter rider detection that provides a significant improvement over the current state of the art, as published in *S. Gilroy, D. Mullins, E. Jones, A. Parsi, and M. Glavin, "E-scooter rider detection and classification in dense urban environments", Results in Engineering, Vol.16, 2022. [14].*

6.3 Future Work

There are a number of future research opportunities available through further development of the concepts described in this thesis.

1. The research conducted in this thesis focuses on partially occluded pedestrians and e-scooter riders in the automotive environment. However, a knowledge gap remains for the provision of an objective metric for cyclist detection that incorporates both the rider and the bicycle. Cyclists follow a very different dynamic profile than pedestrians and early detection and classification is required to ensure safe and efficient path planning for autonomous vehicles in heterogeneous traffic.
2. There is scope for further analysis on the saliency of individual semantic parts on detection confidence for pedestrian detection algorithms and for a com-

prehensive study on the impact of self-occlusion (where parts of a pedestrian overlap or occlude itself due to pedestrian pose) on detectability. This will help to identify scenarios that are currently underrepresented in training datasets and to improve the robustness of future pedestrian detection algorithms.

3. Additional future work will apply the occlusion level classification method described in Chapter 3 to current popular pedestrian detection benchmarks such as the KITTI Vision Benchmark [7], Caltech Pedestrian Detection Benchmark [5], CityPersons Dataset [8] and EuroCity Persons Dataset [9] to provide fine-grained occlusion level annotations and facilitate objective occlusion level analysis using these benchmarks.
4. There is large scope for future work in the field of e-scooter rider detection as this particularly vulnerable class of road user remains largely unrepresented in VRU detection benchmarks. New mobility solutions must be adequately incorporated into future iterations of vulnerable road user detection algorithms as their use becomes more prevalent in society.
5. The method of occlusion level classification and 2D body surface area estimation described in this thesis could be used to improve occlusion-aware pedestrian detection networks through more precise identification of the severity of occlusion and to improve the performance of amodal perception algorithms through more accurate identification of occluded semantic parts and occluded

surface area.

6. Images captured in the automotive environment often contain multiple inhibiting factors to detection such as occlusion, small scale or far away instances, adverse weather and shadows or low light. Although the ideal solution is to have a single detection algorithm for all scenarios, future performance gains may be achieved through the development of more advanced multibranch detection networks. Such networks will first classify the scene for each region of interest and then apply the most appropriate detection algorithm for each scenario, in addition to a priority score based on the relevance of the network for the specific scenario. Although this form of fusion network will incur additional processing costs, if efficiently designed it may yield significant performance improvements that will help to close the gap between the current state of the art and the vulnerable road user detection capabilities required for safe autonomous driving.

Bibliography

- [1] W. H. Organization *et al.*, “Global status report on road safety 2018 (2018),” *Geneva, Switzerland, WHO*, 2019.
- [2] W. H. Organization *et al.*, “Road safety,” 2020.
- [3] S. O. road Automated Vehicles Standards Committee *et al.*, “Sae j3016: Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems,” 2018.
- [4] I. SAE, “Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles j3016_202104 (p. 41),” *SAE International*, 2021.
- [5] P. Dollár, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: A benchmark,” in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 304–311.
- [6] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2011.
- [7] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.
- [8] S. Zhang, R. Benenson, and B. Schiele, “Citypersons: A diverse dataset for pedestrian detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3213–3221.
- [9] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, “Eurocity persons: A novel benchmark for person detection in traffic scenes,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1844–1861, 2019.

-
- [10] S. Gilroy, E. Jones, and M. Glavin, “Overcoming occlusion in the automotive environment—a review,” *IEEE Transactions on Intelligent Transportation Systems*, 2019.
- [11] S. Gilroy, M. Glavin, E. Jones, and D. Mullins, “Pedestrian occlusion level classification using keypoint detection and 2d body surface area estimation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3833–3839.
- [12] S. Gilroy, M. Glavin, E. Jones, and D. Mullins, “An objective method for pedestrian occlusion level classification,” *Pattern Recognition Letters*, vol. 164, pp. 96–103, 2022.
- [13] S. Gilroy, D. Mullins, A. Parsi, E. Jones, and M. Glavin, “Replacing the human driver: An objective benchmark for occluded pedestrian detection,” *Biomimetic Intelligence and Robotics*, vol. 3, no. 3, p. 100115, 2023.
- [14] S. Gilroy, D. Mullins, E. Jones, A. Parsi, and M. Glavin, “E-scooter rider detection and classification in dense urban environments,” *Results in Engineering*, vol. 16, p. 100677, 2022.
- [15] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in *European conference on computer vision*. Springer, 2016, pp. 354–370.
- [16] H. Liu, Y. Agam, J. R. Madsen, and G. Kreiman, “Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex,” *Neuron*, vol. 62, no. 2, pp. 281–290, 2009.
- [17] S. Thorpe, D. Fize, and C. Marlot, “Speed of processing in the human visual system,” *nature*, vol. 381, no. 6582, pp. 520–522, 1996.
- [18] C. P. Heesy, “Seeing in stereo: the ecology and evolution of primate binocular vision and stereopsis,” *Evolutionary Anthropology: Issues, News, and Reviews*, vol. 18, no. 1, pp. 21–35, 2009.
- [19] H. Tang and G. Kreiman, “Recognition of occluded objects,” in *Computational and cognitive neuroscience of vision*. Springer, 2017, pp. 41–58.
- [20] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, “Subcategory-aware convolutional neural networks for object proposals and detection,” in *2017 IEEE winter*

- conference on applications of computer vision (WACV). IEEE, 2017, pp. 924–933.
- [21] X. Hu, X. Xu, Y. Xiao, H. Chen, S. He, J. Qin, and P.-A. Heng, “Sinet: A scale-insensitive convolutional neural network for fast vehicle detection,” *IEEE transactions on intelligent transportation systems*, vol. 20, no. 3, pp. 1010–1019, 2018.
- [22] M. Singh, “Modal and amodal completion generate different shapes,” *Psychological Science*, vol. 15, no. 7, pp. 454–459, 2004.
- [23] D. D. Hoffman and M. Singh, “Saliency of visual parts,” *Cognition*, vol. 63, no. 1, pp. 29–78, 1997.
- [24] K. Fukushima, “Recognition of partly occluded patterns: a neural network model,” *Biological Cybernetics*, vol. 84, no. 4, pp. 251–259, 2001.
- [25] J. S. Johnson and B. A. Olshausen, “The recognition of partially visible natural objects in the presence and absence of their occluders,” *Vision research*, vol. 45, no. 25-26, pp. 3262–3276, 2005.
- [26] M. Meng and M. C. Potter, “Detecting and remembering pictures with and without visual noise,” *Journal of Vision*, vol. 8, no. 9, pp. 7–7, 2008.
- [27] M. Struwe, “Active occlusion-handling for appearance-based object recognition models,” Ph.D. dissertation, Dissertation, Bielefeld, Universität Bielefeld, 2016, 2017.
- [28] A. M. Fyall, Y. El-Shamayleh, H. Choi, E. Shea-Brown, and A. Pasupathy, “Dynamic representation of partially occluded objects in primate prefrontal and visual cortex,” *Elife*, vol. 6, p. e25784, 2017.
- [29] T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio, “A quantitative theory of immediate visual recognition,” *Progress in brain research*, vol. 165, pp. 33–56, 2007.
- [30] T. Masquelier and S. J. Thorpe, “Unsupervised learning of visual features through spike timing dependent plasticity,” *PLoS computational biology*, vol. 3, no. 2, p. e31, 2007.
- [31] S.-M. Khaligh-Razavi and N. Kriegeskorte, “Deep supervised, but not unsupervised, models may explain it cortical representation,” *PLoS computational biology*, vol. 10, no. 11, p. e1003915, 2014.

-
- [32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012.
- [33] U. Güçlü and M. A. van Gerven, “Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream,” *Journal of Neuroscience*, vol. 35, no. 27, pp. 10 005–10 014, 2015.
- [34] D. Wyatte, T. Curran, and R. O’Reilly, “The limits of feedforward vision: Recurrent processing promotes robust object recognition when objects are degraded,” *Journal of Cognitive Neuroscience*, vol. 24, no. 11, pp. 2248–2261, 2012.
- [35] R. C. O’Reilly, D. Wyatte, S. Herd, B. Mingus, and D. J. Jilk, “Recurrent processing during object recognition,” *Frontiers in psychology*, vol. 4, p. 124, 2013.
- [36] C. J. Spoeer, P. McClure, and N. Kriegeskorte, “Recurrent convolutional neural networks: a better model of biological object recognition,” *Frontiers in psychology*, vol. 8, p. 1551, 2017.
- [37] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, “Towards reaching human performance in pedestrian detection,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 973–986, 2017.
- [38] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, “How far are we from solving pedestrian detection?” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1259–1267.
- [39] D. Hoiem, A. N. Stein, A. A. Efros, and M. Hebert, “Recovering occlusion boundaries from a single image,” in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.
- [40] D. Seychell and C. J. Debono, “Efficient object selection using depth and texture information,” in *2016 Visual Communications and Image Processing (VCIP)*. IEEE, 2016, pp. 1–4.
- [41] D. Hoiem, A. A. Efros, and M. Hebert, “Recovering surface layout from an image,” *International Journal of Computer Vision*, vol. 75, no. 1, pp. 151–172, 2007.

-
- [42] B. Y. Lee, L. H. Liew, W. S. Cheah, and Y. C. Wang, “Occlusion handling in videos object tracking: A survey,” in *IOP conference series: earth and environmental science*, vol. 18, no. 1. IOP Publishing, 2014, p. 012020.
- [43] X. Chen, Q. Li, D. Zhao, and Q. Zhao, “Occlusion cues for image scene layering,” *Computer Vision and Image Understanding*, vol. 117, no. 1, pp. 42–55, 2013.
- [44] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, “Multi-cue pedestrian classification with partial occlusion handling,” in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 990–997.
- [45] G. Palou and P. Salembier, “Monocular depth ordering using t-junctions and convexity occlusion cues,” *IEEE transactions on image processing*, vol. 22, no. 5, pp. 1926–1939, 2013.
- [46] B. Rezaeirowshan, C. Ballester, and G. Haro Ortega, “Monocular depth ordering using perceptual occlusion cues,” in *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)-Volume 4; 2016 Feb 27-29; Rome, Italy. Setúbal: Scitepress; 2016*. SCITEPRESS–Science and Technology Publications, Lda., 2016.
- [47] A. Guzman-Arenas, “Computer recognition of three-dimensional objects in a visual scene.” Massachusetts Institute of Technology Cambridge Project Mac, Tech. Rep., 1968.
- [48] J. McDermott, “Psychophysics with junctions in real images,” *Perception*, vol. 33, no. 9, pp. 1101–1127, 2004.
- [49] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei, “Imagenet large scale visual recognition competition 2012 (ilsvrc2012),” *See net.org/challenges/LSVRC*, vol. 41, p. 6, 2012.
- [50] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” *Proceedings of the IEEE*, 2023.
- [51] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari, “The history began from alexnet: A comprehensive survey on deep learning approaches,” *arXiv preprint arXiv:1803.01164*, 2018.

-
- [52] W. Chen, Y. Zhu, Z. Tian, F. Zhang, and M. Yao, “Occlusion and multi-scale pedestrian detection a review,” *Array*, p. 100318, 2023.
- [53] J. Qi, Y. Gao, X. Liu, Y. Hu, X. Wang, X. Bai, P. H. Torr, S. Belongie, A. Yuille, and S. Bai, “Occluded video instance segmentation,” *arXiv preprint arXiv:2102.01558*, 2021.
- [54] X. Wang, T. X. Han, and S. Yan, “An hog-lbp human detector with partial occlusion handling,” in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 32–39.
- [55] R. M. op het Veld, R. Wijnhoven, Y. Bondarev *et al.*, “Detection and handling of occlusion in an object detection system,” in *Video Surveillance and Transportation Imaging Applications 2015*, vol. 9407. SPIE, 2015, pp. 184–195.
- [56] T. Gao, B. Packer, and D. Koller, “A segmentation-aware object detection model with occlusion handling,” in *CVPR 2011*. IEEE Computer Society, 2011, pp. 1361–1368.
- [57] Z. Chen, T. Xu, and Z. Han, “Occluded face recognition based on the improved svm and block weighted lbp,” in *2011 International Conference on Image Analysis and Signal Processing*. IEEE, 2011, pp. 118–122.
- [58] D. T. Nguyen, W. Li, and P. O. Ogunbona, “Inter-occlusion reasoning for human detection based on variational mean field,” *Neurocomputing*, vol. 110, pp. 51–61, 2013.
- [59] K. C. Chan, A. Ayvaci, and B. Heisele, “Partially occluded object detection by finding the visible features and parts,” in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 2130–2134.
- [60] W. Ouyang and X. Wang, “A discriminative deep model for pedestrian detection with occlusion handling,” in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3258–3265.
- [61] W. Ouyang and X. Wang, “Joint deep learning for pedestrian detection,” in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2056–2063.
- [62] W. Ouyang, X. Zeng, and X. Wang, “Partial occlusion handling in pedestrian detection with a deep model,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 2123–2137, 2015.

-
- [63] T. Baumgartner, D. Mitzel, and B. Leibe, “Tracking people and their objects,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3658–3665.
- [64] S. Kwak, W. Nam, B. Han, and J. H. Han, “Learning occlusion with likelihoods for visual tracking,” in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1551–1558.
- [65] C. Zhou and J. Yuan, “Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection,” in *Asian Conference on Computer Vision*. Springer, 2016, pp. 305–320.
- [66] C. Wojek, S. Walk, S. Roth, and B. Schiele, “Monocular 3d scene understanding with explicit occlusion reasoning,” in *CVPR 2011*. IEEE, 2011, pp. 1993–2000.
- [67] C. Zhou and J. Yuan, “Multi-label learning of part detectors for heavily occluded pedestrian detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3486–3495.
- [68] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool, “Handling occlusions with franken-classifiers,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1505–1512.
- [69] S. Tang, M. Andriluka, and B. Schiele, “Detection and tracking of occluded people,” *International Journal of Computer Vision*, vol. 110, no. 1, pp. 58–69, 2014.
- [70] W. Ouyang and X. Wang, “Single-pedestrian detection aided by multi-pedestrian detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3198–3205.
- [71] R. Rosales and S. Sclaroff, “Improved tracking of multiple humans with trajectory prediction and occlusion modeling,” Boston University Computer Science Department, Tech. Rep., 1998.
- [72] P. F. Gabriel, J. G. Verly, J. H. Piater, and A. Genon, “The state of the art in multiple object tracking under occlusion in video sequences,” in *Advanced Concepts for Intelligent Vision Systems*. Citeseer, 2003, pp. 166–173.

-
- [73] J. Zhao, W. Qiao, and G.-Z. Men, “An approach based on mean shift and kalman filter for target tracking under occlusion,” in *2009 International Conference on Machine Learning and Cybernetics*, vol. 4. IEEE, 2009, pp. 2058–2062.
- [74] A. Ali and K. Terada, “A framework for human tracking using kalman filter and fast mean shift algorithms,” in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, 2009, pp. 1028–1033.
- [75] S. Vethamani and K. Diala, “Spatio-temporal approaches for handling occlusions based on object tracking,” *J. Comput. Sci. Syst. Biol.*, vol. 10, no. 4, pp. 93–97, 2017.
- [76] A. Sadeghian, A. Alahi, and S. Savarese, “Tracking the untrackable: Learning to track multiple cues with long-term dependencies,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 300–311.
- [77] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, “Visual tracking: An experimental survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1442–1468, 2013.
- [78] Z. Kalal, J. Matas, and K. Mikolajczyk, “Online learning of robust object detectors during unstable tracking,” in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, 2009, pp. 1417–1424.
- [79] M. Ozuysal, P. Fua, and V. Lepetit, “Fast keypoint recognition in ten lines of code,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2007, pp. 1–8.
- [80] M. Fiaz, A. Mahmood, and S. K. Jung, “Tracking noisy targets: A review of recent object tracking approaches,” *arXiv preprint arXiv:1802.03098*, 2018.
- [81] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, “Eco: Efficient convolution operators for tracking,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6638–6646.
- [82] A. Lukezic, T. Vojir, L. ˇCehovin Zajc, J. Matas, and M. Kristan, “Discriminative correlation filter with channel and spatial reliability,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6309–6318.

-
- [83] S. K. Kwon, E. Hyun, J.-H. Lee, J. Lee, and S. H. Son, “A low-complexity scheme for partially occluded pedestrian detection using lidar-radar sensor fusion,” in *2016 IEEE 22nd International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*. IEEE, 2016, pp. 104–104.
- [84] S. K. Kwon, E. Hyun, J.-H. Lee, J. Lee, and S. H. Son, “Detection scheme for a partially occluded pedestrian based on occluded depth in lidar–radar sensor fusion,” *Optical Engineering*, vol. 56, no. 11, p. 113112, 2017.
- [85] Y. Kim, S. Ha, and J. Kwon, “Human detection using doppler radar based on physical characteristics of targets,” *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 2, pp. 289–293, 2014.
- [86] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, “Frustum pointnets for 3d object detection from rgb-d data,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.
- [87] T.-E. Wu, C.-C. Tsai, and J.-I. Guo, “Lidar/camera sensor fusion technology for pedestrian detection,” in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1675–1678.
- [88] S. Schneider, M. Himmelsbach, T. Luettel, and H.-J. Wuensche, “Fusing vision and lidar-synchronization, correction and occlusion reasoning,” in *2010 IEEE Intelligent Vehicles Symposium*. IEEE, 2010, pp. 388–393.
- [89] C. Premebida and U. Nunes, “Fusing lidar, camera and semantic information: A context-based approach for pedestrian detection,” *The International Journal of Robotics Research*, vol. 32, no. 3, pp. 371–384, 2013.
- [90] C. Premebida, J. Carreira, J. Batista, and U. Nunes, “Pedestrian detection combining rgb and dense lidar data,” in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 4112–4117.
- [91] F. García, J. García, A. Ponz, A. De La Escalera, and J. M. Armingol, “Context aided pedestrian detection for danger estimation based on laser scanner and computer vision,” *Expert Systems with Applications*, vol. 41, no. 15, pp. 6646–6661, 2014.

-
- [92] C.-M. Huang and B.-W. Jiang, “Occlusion handling of visual tracking by fusing multiple visual clues,” in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017, pp. 836–839.
- [93] Y.-C. Liu, S.-S. Huang, C.-H. Lu, F.-C. Chang, and P.-Y. Lin, “Thermal pedestrian detection using block lbp with multi-level classifier,” in *2017 International Conference on Applied System Innovation (ICASI)*. IEEE, 2017, pp. 602–605.
- [94] N. B. Bo, P. Veelaert, and W. Philips, “Occlusion robust symbol level fusion for multiple people tracking.” in *VISIGRAPP (6: VISAPP)*, 2017, pp. 216–226.
- [95] B. Tang, S. Chien, Z. Huang, and Y. Chen, “Pedestrian protection using the integration of v2v and the pedestrian automatic emergency braking system,” 2016.
- [96] S. Y. Gelbal, S. Arslan, H. Wang, B. Aksun-Guvenc, and L. Guvenc, “Elastic band based pedestrian collision avoidance using v2x communication,” in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 270–276.
- [97] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge 2012 (voc2012) results. 2012 <http://www.pascal-network.org/challenges>,” in *VOC/voc2012/workshop/index.html*, 2012.
- [98] F. Yang, W. Choi, and Y. Lin, “Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2129–2137.
- [99] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [100] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu, “Accurate single stage detector using recurrent rolling convolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5420–5428.
- [101] S.-I. Jung and K.-S. Hong, “Deep network aided by guiding network for pedestrian detection,” *Pattern Recognition Letters*, vol. 90, pp. 43–49, 2017.

- [102] 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 Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.
- [103] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, “Mot16: A benchmark for multi-object tracking,” *arXiv preprint arXiv:1603.00831*, 2016.
- [104] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [105] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, “Multispectral pedestrian detection: Benchmark dataset and baseline,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.
- [106] G. Brazil, X. Yin, and X. Liu, “Illuminating pedestrians via simultaneous detection & segmentation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4950–4959.
- [107] X. Du, M. El-Khamy, J. Lee, and L. Davis, “Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection,” in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 953–961.
- [108] S. Wang, J. Cheng, H. Liu, and M. Tang, “Pcn: Part and context information for pedestrian detection with cnns,” *arXiv preprint arXiv:1804.04483*, 2018.
- [109] Y. Tian, P. Luo, X. Wang, and X. Tang, “Deep learning strong parts for pedestrian detection,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1904–1912.
- [110] A. Ewert, M. Brost, and S. Schmid, “Small electric vehicles—benefits and drawbacks for sustainable urban development,” *Small Electric Vehicles*, p. 3, 2021.
- [111] C. Hardt and K. Bogenberger, “Usage of e-scooters in urban environments,” *Transportation research procedia*, vol. 37, pp. 155–162, 2019.
- [112] E. Parliament and of the Council of the European Union, “Regulation (eu) no 168/2013 of the european parliament and of the council of 15 january 2013

- on the approval and market surveillance of two-or three-wheel vehicles and quadricycles,” 2013.
- [113] Y. Pang, J. Cao, Y. Li, J. Xie, H. Sun, and J. Gong, “Tju-dhd: A diverse high-resolution dataset for object detection,” *IEEE Transactions on Image Processing*, vol. 30, pp. 207–219, 2020.
- [114] X. Li, F. Flohr, Y. Yang, H. Xiong, M. Braun, S. Pan, K. Li, and D. M. Gavrila, “A new benchmark for vision-based cyclist detection,” in *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 1028–1033.
- [115] X. Li, L. Li, F. Flohr, J. Wang, H. Xiong, M. Bernhard, S. Pan, D. M. Gavrila, and K. Li, “A unified framework for concurrent pedestrian and cyclist detection,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 2, pp. 269–281, 2017.
- [116] Y.-T. Hu, H.-S. Chen, K. Hui, J.-B. Huang, and A. G. Schwing, “Sail-vos: Semantic amodal instance level video object segmentation—a synthetic dataset and baselines,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3105–3115.
- [117] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, “Crowd-human: A benchmark for detecting human in a crowd,” *arXiv preprint arXiv:1805.00123*, 2018.
- [118] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, “Pedhunter: Occlusion robust pedestrian detector in crowded scenes,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 639–10 646.
- [119] P. Chaudhary, S. D’Aronco, M. Moy de Vitry, J. P. Leitão, and J. D. Wegner, “Flood-water level estimation from social media images,” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, no. 2/W5, pp. 5–12, 2019.
- [120] Y. Feng, C. Brenner, and M. Sester, “Flood severity mapping from volunteered geographic information by interpreting water level from images containing people: A case study of hurricane harvey,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 301–319, 2020.
- [121] K.-A. C. Quan, V.-T. Nguyen, T.-C. Nguyen, T. V. Nguyen, and M.-T. Tran, “Flood level prediction via human pose estimation from social media images,”

- in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 479–485.
- [122] J. Noh, S. Lee, B. Kim, and G. Kim, “Improving occlusion and hard negative handling for single-stage pedestrian detectors,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 966–974.
- [123] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, “Occlusion-aware r-cnn: Detecting pedestrians in a crowd,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 637–653.
- [124] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.
- [125] A. Wallace, “The exposure treatment of burns,” *The Lancet*, vol. 257, no. 6653, pp. 501–504, 1951.
- [126] G. A. Knaysi, G. F. Crikelair, and B. Cosman, “The rule of nines: its history and accuracy,” *Plastic and reconstructive surgery*, vol. 41, no. 6, pp. 560–563, 1968.
- [127] K. Borhani-Khomani, S. Partoft, and R. Holmgaard, “Assessment of burn size in obese adults; a literature review,” *Journal of plastic surgery and hand surgery*, vol. 51, no. 6, pp. 375–380, 2017.
- [128] I. Tocco-Tussardi, B. Presman, and F. Huss, “Want correct percentage of tbsa burned? let a layman do the assessment,” *Journal of Burn Care & Research*, vol. 39, no. 2, pp. 295–301, 2018.
- [129] E. H. Livingston and S. Lee, “Percentage of burned body surface area determination in obese and nonobese patients,” *Journal of surgical research*, vol. 91, no. 2, pp. 106–110, 2000.
- [130] J. Yin and G. Sun, “Adaptive multi-strategy for multi-vehicle with mutual occlusion tracking,” in *2014 International Conference on Audio, Language and Image Processing*. IEEE, 2014, pp. 743–748.
- [131] R. Velazquez-Pupo, A. Sierra-Romero, D. Torres-Roman, Y. V. Shkvarko, J. Santiago-Paz, D. Gómez-Gutiérrez, D. Robles-Valdez, F. Hermosillo-Reynoso, and M. Romero-Delgado, “Vehicle detection with occlusion handling,

- tracking, and oc-svm classification: A high performance vision-based system,” *Sensors*, vol. 18, no. 2, p. 374, 2018.
- [132] W. Zhang, Q. J. Wu, X. Yang, and X. Fang, “Multilevel framework to detect and handle vehicle occlusion,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 1, pp. 161–174, 2008.
- [133] A. Ghasemi and R. Safabakhsh, “A real-time multiple vehicle classification and tracking system with occlusion handling,” in *2012 IEEE 8th International Conference on Intelligent Computer Communication and Processing*. IEEE, 2012, pp. 109–115.
- [134] W. Fang, Y. Zhao, Y. Yuan, and K. Liu, “Real-time multiple vehicles tracking with occlusion handling,” in *2011 Sixth International Conference on Image and Graphics*. IEEE, 2011, pp. 667–672.
- [135] L. Huang and M. Barth, “Real-time multi-vehicle tracking based on feature detection and color probability model,” in *2010 IEEE Intelligent Vehicles Symposium*. IEEE, 2010, pp. 981–986.
- [136] T. Zhang, K. Jia, C. Xu, Y. Ma, and N. Ahuja, “Partial occlusion handling for visual tracking via robust part matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1258–1265.
- [137] E. Galceran, E. Olson, and R. M. Eustice, “Augmented vehicle tracking under occlusions for decision-making in autonomous driving,” in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 3559–3565.
- [138] W. Min, M. Fan, X. Guo, and Q. Han, “A new approach to track multiple vehicles with the combination of robust detection and two classifiers,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 174–186, 2017.
- [139] H. Van Pham and B.-R. Lee, “Front-view car detection and counting with occlusion in dense traffic flow,” *International Journal of Control, Automation and Systems*, vol. 13, no. 5, pp. 1150–1160, 2015.
- [140] E. Ohn-Bar, S. Sivaraman, and M. Trivedi, “Partially occluded vehicle recognition and tracking in 3d,” in *2013 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2013, pp. 1350–1355.

-
- [141] J. Wang, C. Xie, Z. Zhang, J. Zhu, L. Xie, and A. Yuille, “Detecting semantic parts on partially occluded objects,” *arXiv preprint arXiv:1707.07819*, 2017.
- [142] J. Li, H.-C. Wong, S.-L. Lo, and Y. Xin, “Multiple object detection by a deformable part-based model and an r-cnn,” *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 288–292, 2018.
- [143] H. T. Niknejad, A. Takeuchi, S. Mita, and D. McAllester, “On-road multivehicle tracking using deformable object model and particle filter with improved likelihood estimation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 748–758, 2012.
- [144] H. T. Niknejad, T. Kawano, Y. Oishi, and S. Mita, “Occlusion handling using discriminative model of trained part templates and conditional random field,” in *2013 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2013, pp. 750–755.
- [145] C. Wang, Y. Fang, H. Zhao, C. Guo, S. Mita, and H. Zha, “Probabilistic inference for occluded and multiview on-road vehicle detection,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 215–229, 2015.
- [146] S. Sivaraman and M. M. Trivedi, “Vehicle detection by independent parts for urban driver assistance,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1597–1608, 2013.
- [147] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, “Data-driven 3d voxel patterns for object category recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1903–1911.
- [148] B. Li, W. Hu, T. Wu, and S.-C. Zhu, “Modeling occlusion by discriminative and-or structures,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2560–2567.
- [149] B. Li, T. Wu, and S.-C. Zhu, “Integrating context and occlusion for car detection by hierarchical and-or model,” in *European Conference on Computer Vision*. Springer, 2014, pp. 652–667.
- [150] T. Wu, B. Li, and S.-C. Zhu, “Learning and-or model to represent context and occlusion for car detection and viewpoint estimation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1829–1843, 2015.

-
- [151] W. Chu, Y. Liu, C. Shen, D. Cai, and X.-S. Hua, "Multi-task vehicle detection with region-of-interest voting," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 432–441, 2017.
- [152] J. Chung and K. Sohn, "Image-based learning to measure traffic density using a deep convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1670–1675, 2017.
- [153] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teuliere, and T. Chateau, "Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2040–2049.
- [154] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Craft objects from images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 6043–6051.
- [155] H. N. Phan, L. H. Pham, D. N.-N. Tran, and S. V.-U. Ha, "Occlusion vehicle detection algorithm in crowded scene for traffic surveillance system," in *2017 International Conference on System Science and Engineering (ICSSE)*. IEEE, 2017, pp. 215–220.
- [156] C.-C. Chiu, M.-Y. Ku, and H.-T. Chen, "Motorcycle detection and tracking system with occlusion segmentation," in *Eighth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'07)*. IEEE, 2007, pp. 32–32.
- [157] M.-Y. Ku, C.-C. Chiu, H.-T. Chen, and S.-H. Hong, "Visual motorcycle detection and tracking algorithms," *WSEAS Trans. Electron*, vol. 5, no. 4, pp. 121–131, 2008.
- [158] J. Chiverton, "Helmet presence classification with motorcycle detection and tracking," *IET Intelligent Transport Systems*, vol. 6, no. 3, pp. 259–269, 2012.
- [159] J.-C. Tai and K.-T. Song, "Image tracking of motorcycles and vehicles on urban roads and its application to traffic monitoring and enforcement," *Journal of the Chinese Institute of Engineers*, vol. 33, no. 6, pp. 923–933, 2010.
- [160] M. Ashvini, G. Revathi, B. Yogameena, and S. Saravanaperumaal, "View invariant motorcycle detection for helmet wear analysis in intelligent traffic surveillance," in *Proceedings of International Conference on Computer Vision and Image Processing*. Springer, 2017, pp. 175–185.

- [161] W. Tian and M. Lauer, “Fast and robust cyclist detection for monocular camera systems,” in *International joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, 2015.
- [162] X.-P. Du, H. Xiong, and Y. Li, “Have a deep look at deformable part models for cyclist detection,” in *Computer Science and Artificial Intelligence: Proceedings of the International Conference on Computer Science and Artificial Intelligence (CSAI2016)*. World Scientific, 2018, pp. 147–154.
- [163] J. J. Anaya, E. Talavera, D. Giménez, N. Gómez, F. Jiménez, and J. E. Naranjo, “Vulnerable road users detection using v2x communications,” in *2015 IEEE 18th international conference on intelligent transportation systems*. IEEE, 2015, pp. 107–112.
- [164] J. J. Anaya, A. Ponz, F. García, and E. Talavera, “Motorcycle detection for adas through camera and v2v communication, a comparative analysis of two modern technologies,” *Expert Systems with Applications*, vol. 77, pp. 148–159, 2017.
- [165] J. E. Naranjo, F. Jiménez, J. J. Anaya, E. Talavera, and O. Gómez, “Application of vehicle to another entity (v2x) communications for motorcycle crash avoidance,” *Journal of Intelligent Transportation Systems*, vol. 21, no. 4, pp. 285–295, 2017.
- [166] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, “3d bounding box estimation using deep learning and geometry,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 7074–7082.
- [167] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- [168] K. Apurv, R. Tian, and R. Sherony, “Detection of e-scooter riders in naturalistic scenes,” *arXiv preprint arXiv:2111.14060*, 2021.
- [169] H. Nguyen, M. Nguyen, and Q. Sun, “Electric scooter and its rider detection framework based on deep learning for supporting scooter-related injury emergency services,” in *International Symposium on Geometry and Vision*. Springer, 2021, pp. 233–246.

-
- [170] C. P. S. University, “E-scooter counting on sidewalks with machine learning,” *Online*. <https://dxhub.calpoly.edu/challenges/escooter-counting-on-sidewalks/> (accessed: 26th April 2022), 2022.
- [171] D. C. P. S. University, “Santa monica e-scooter detection,” *GitHub*. <https://github.com/cal-poly-dxhub/Santa-Monica-Scooter-Detection> (accessed: 26th April 2022), 2022.
- [172] J. L. Chu and A. Krzyżak, “The recognition of partially occluded objects with support vector machines, convolutional neural networks and deep belief networks,” *Journal of Artificial Intelligence and Soft Computing Research*, vol. 4, no. 1, pp. 5–19, 2014.
- [173] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, “On deep generative models with applications to recognition,” in *CVPR 2011*. IEEE, 2011, pp. 2857–2864.
- [174] A. Cavagna, S. Melillo, L. Parisi, and F. Ricci-Tersenghi, “Sparta-tracking across occlusions via global partitioning of 3d clouds of points,” *arXiv preprint arXiv:1802.05878*, 2018.
- [175] C. Liu, F. Chang, Z. Chen, and D. Liu, “Fast traffic sign recognition via high-contrast region extraction and extended sparse representation,” *IEEE transactions on Intelligent transportation systems*, vol. 17, no. 1, pp. 79–92, 2015.
- [176] P. Huang, M. Cheng, Y. Chen, H. Luo, C. Wang, and J. Li, “Traffic sign occlusion detection using mobile laser scanning point clouds,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2364–2376, 2017.
- [177] V. A. Prisacariu, R. Timofte, K. Zimmermann, I. Reid, and L. Van Gool, “Integrating object detection with 3d tracking towards a better driver assistance system,” in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3344–3347.
- [178] Y. Xie, L.-f. Liu, C.-h. Li, and Y.-y. Qu, “Unifying visual saliency with hog feature learning for traffic sign detection,” in *2009 IEEE Intelligent Vehicles Symposium*. IEEE, 2009, pp. 24–29.
- [179] L. Li, J. Li, and J. Sun, “Robust traffic sign detection using fuzzy shape recognizer,” in *MIPPR 2009: Pattern Recognition and Computer Vision*, vol. 7496. SPIE, 2009, pp. 269–276.

-
- [180] B. Soheilian, N. Paparoditis, and B. Vallet, “Detection and 3d reconstruction of traffic signs from multiple view color images,” *ISPRS journal of photogrammetry and remote sensing*, vol. 77, pp. 1–20, 2013.
- [181] T. Zhang, Q. Ye, B. Zhang, J. Liu, X. Zhang, and Q. Tian, “Feature calibration network for occluded pedestrian detection,” *IEEE Transactions on Intelligent Transportation Systems*, 2020.
- [182] B. Ruan and C. Zhang, “Occluded pedestrian detection combined with semantic features,” *IET Image Processing*, 2021.
- [183] E. Vejbjørn, S. Mats, P. Angelo, B. Gisela, J. Sebastian, W. Johan, H. Alena *et al.*, “The illusion of absence: how a common feature of magic shows can explain a class of road accidents,” *Cognitive Research*, vol. 6, no. 1, 2021.
- [184] C. Ning, L. Menglu, Y. Hao, S. Xueping, and L. Yunhong, “Survey of pedestrian detection with occlusion,” *Complex & Intelligent Systems*, vol. 7, no. 1, pp. 577–587, 2021.
- [185] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, “From handcrafted to deep features for pedestrian detection: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [186] Y. Xiao, K. Zhou, G. Cui, L. Jia, Z. Fang, X. Yang, and Q. Xia, “Deep learning for occluded and multi-scale pedestrian detection: A review,” *IET Image Processing*, vol. 15, no. 2, pp. 286–301, 2021.
- [187] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, “Generalizable pedestrian detection: The elephant in the room,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 328–11 337.
- [188] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, “Detectron2,” <https://github.com/facebookresearch/detectron2>, 2019.
- [189] 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,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [190] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

-
- [191] J. Zhuo, Z. Chen, J. Lai, and G. Wang, “Occluded person re-identification,” in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.
- [192] J. Marín, D. Vázquez, A. M. López, J. Amores, and L. I. Kuncheva, “Occlusion handling via random subspace classifiers for human detection,” *IEEE transactions on cybernetics*, vol. 44, no. 3, pp. 342–354, 2013.
- [193] Q. Zhou, S. Wang, Y. Wang, Z. Huang, and X. Wang, “Human de-occlusion: Invisible perception and recovery for humans,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3691–3701.
- [194] X. Li, L. Li, F. Flohr, J. Wang, H. Xiong, M. Bernhard, S. Pan, D. M. Gavrila, and K. Li, “A unified framework for concurrent pedestrian and cyclist detection,” *IEEE transactions on intelligent transportation systems*, vol. 18, no. 2, pp. 269–281, 2016.
- [195] S. Walk, N. Majer, K. Schindler, and B. Schiele, “New features and insights for pedestrian detection,” in *2010 IEEE Computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 1030–1037.
- [196] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, “An exploration of why and when pedestrian detection fails,” in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, 2015, pp. 2335–2340.
- [197] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, “What can help pedestrian detection?” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3127–3136.
- [198] N. Ragesh and R. Rajesh, “Pedestrian detection in automotive safety: understanding state-of-the-art,” *IEEE Access*, vol. 7, pp. 47 864–47 890, 2019.
- [199] J. Cao, Y. Pang, J. Han, B. Gao, and X. Li, “Taking a look at small-scale pedestrians and occluded pedestrians,” *IEEE transactions on image processing*, vol. 29, pp. 3143–3152, 2019.
- [200] T. Toprak, B. A. Can, M. Ozcelikors, S. B. Tekin, and M. A. Selver, “Limitations of feature-classifier strategies on pedestrian detection for self driving cars,” in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2020, pp. 1–6.

-
- [201] H. Chandel and S. Vatta, “Occlusion detection and handling: a review,” *International Journal of Computer Applications*, vol. 120, no. 10, 2015.
- [202] K. Saleh, S. Szénási, and Z. Vámosy, “Occlusion handling in generic object detection: A review,” in *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMII)*. IEEE, 2021, pp. 000 477–000 484.
- [203] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, “Partial person re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4678–4686.
- [204] Pytorch, “Torchvision model zoo,” *Online*. <https://pytorch.org/vision/stable/models.html>, 2022.
- [205] B. E. Moore and J. J. Corso, “Fiftyone model zoo,” *Online*. https://voxel51.com/docs/fiftyone/user_guide/model_zoo, 2022.
- [206] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: towards real-time object detection with region proposal networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.
- [207] J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: Object detection via region-based fully convolutional networks,” in *Advances in neural information processing systems*, 2016, pp. 379–387.
- [208] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision*. Springer, 2016, pp. 21–37.
- [209] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, “Searching for mobilenetv3,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.
- [210] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [211] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

-
- [212] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” *arXiv preprint arXiv:1904.07850*, 2019.
- [213] L. Chen, S. Lin, X. Lu, D. Cao, H. Wu, C. Guo, C. Liu, and F.-Y. Wang, “Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey,” *IEEE Transactions on Intelligent Transportation Systems*, 2021.
- [214] B. E. Moore and J. J. Corso, “Fiftyone,” *GitHub. Note: <https://github.com/voxel51/fiftyone>*, 2020.
- [215] K. Heineke, B. Kloss, D. Scurtu, and F. Weig, “Micromobility’s 15,000-mile checkup,” *Retrieved from McKinsey & Company Automotive & Assembly: <https://www.mckinsey.com/industries/automotive-andassembly/our-insights/micromobilitys-15000-mile-checkup>*, 2019.
- [216] K. L. Ioannides, P.-C. Wang, K. Kowsari, V. Vu, N. Kojima, D. Clayton, C. Liu, T. K. Trivedi, D. L. Schriger, and J. G. Elmore, “E-scooter related injuries: Using natural language processing to rapidly search 36 million medical notes,” *PloS one*, vol. 17, no. 4, p. e0266097, 2022.
- [217] L. F. Beck, A. M. Dellinger, and M. E. O’neil, “Motor vehicle crash injury rates by mode of travel, united states: using exposure-based methods to quantify differences,” *American Journal of Epidemiology*, vol. 166, no. 2, pp. 212–218, 2007.
- [218] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.
- [219] A. Krizhevsky, “One weird trick for parallelizing convolutional neural networks,” *arXiv preprint arXiv:1404.5997*, 2014.
- [220] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size,” *arXiv preprint arXiv:1602.07360*, 2016.
- [221] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [222] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

- [223] J. Howard and S. Gugger, “Fastai: a layered api for deep learning,” *Information*, vol. 11, no. 2, p. 108, 2020.