

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

Title	High-accuracy facial depth models derived from 3D synthetic data
Author(s)	Khan, Faisal; Basak, Shubhajit; Javidnia, Hossein; Schukat, Michael; Corcoran, Peter
Publication Date	2020-08-31
Publication Information	Khan, Faisal, Basak, Shubhajit, Javidnia, Hossein, Schukat, Michael, & Corcoran, Peter. (2020). High-accuracy facial depth models derived from 3D synthetic data. Paper presented at the 31st Irish Signals and Systems Conference (ISSC), Letterkenny, Ireland, 11-12 June, DOI: 10.1109/ISSC49989.2020.9180166.
Publisher	Institute of Electrical and Electronics Engineers (IEEE)
Link to publisher's version	<a href="https://dx.doi.org/10.1109/ISSC49989.2020.9180166">https://dx.doi.org/10.1109/ISSC49989.2020.9180166</a>
Item record	<a href="http://hdl.handle.net/10379/16790">http://hdl.handle.net/10379/16790</a>
DOI	<a href="http://dx.doi.org/10.1109/ISSC49989.2020.9180166">http://dx.doi.org/10.1109/ISSC49989.2020.9180166</a>

Downloaded 2024-04-24T22:26:10Z

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



# High-Accuracy Facial Depth Models derived from 3D Synthetic Data

Faisal Khan  
College of Engineering and Informatics  
National University Ireland Galway  
Galway, Ireland  
f.khan4@nuigalway.ie

Michael Schukat  
College of Engineering and Informatics  
National University Ireland Galway  
Galway, Ireland  
micheal.schukat@nuigalway.ie

Shubhajit Basak  
College of Engineering and Informatics  
National University Ireland Galway  
Galway, Ireland  
s.basak1@nuigalway.ie

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

Hossein Javidnia  
ADAPT Center, O Reilly Institute  
Trinity College Dublin Ireland  
25 Westland Row, Dublin, 2, Ireland  
hossein.javidnia@adaptcentre.ie

**Abstract**—In this paper, we explore how synthetically generated 3D face models can be used to construct a high-accuracy ground truth for depth. This allows us to train the Convolutional Neural Networks (CNN) to solve facial depth estimation problems. These models provide sophisticated controls over image variations including pose, illumination, facial expressions and camera position. 2D training samples can be rendered from these models, typically in RGB format, together with depth information. Using synthetic facial animations, a dynamic facial expression or facial action data can be rendered for a sequence of image frames together with ground truth depth and additional metadata such as head pose, light direction, etc. The synthetic data is used to train a CNN-based facial depth estimation system which is validated on both synthetic and real images. Potential fields of application include 3D reconstruction, driver monitoring systems, robotic vision systems, and advanced scene understanding.

**Keywords**—3D Facial models, Facial depth, Face attributes, Facial image dataset

## I. INTRODUCTION

Estimating human shape, pose, motion and depth from images are fundamental challenges for many multimedia applications and provide information that can be leveraged to enhance quality and immersion in advanced consumer use cases. Examples include scene analysis & understanding, human behaviour analysis, driver monitoring for semi-autonomous driving, augmented reality systems and facial expression analysis and facial authentication. Today, state-of-art systems for these use cases will rely on highly optimized convolutional neural networks designed to run on low-power embedded hardware. Such solutions require large, high-quality training datasets.

Facial images, in particular, are at the core of many consumer multimedia systems. They exhibit rich variations in pose, hairstyle, expression, structure and their 2D appearance is affected by external factors such as lighting and camera location. Many face variations can be synthesized using existing advanced 3D tools such as iClone [1] and Blender [2]. Using these tools, it is feasible to generate a large number of synthetic images required for training Convolutional Neural Network (CNN) models. Rendering synthetic facial images would be highly useful for numerous tasks as it can provide enough realism to create various ground truth in terms of occlusions, depth, motion, body-part segmentation, camera and light direction.

The current generation of deep learning models requires the datasets to contain various information and accurate data for the training and evaluation process. The existing human facial datasets do not have the accurate depth information that defines the actual position of each facial element. The depth information in these datasets requires the manual description of the scene, which is an error-prone and time-consuming task especially dealing with video [3]. In such type of facial dataset, they are not sufficiently large and varied enough to learn the CNN models, as a consequence, they come with a low performance which restricts real-world applications [4-5].

Recently deep learning-based methodologies have significantly improved the performances of face recognition systems, Human-Computer Interaction (HCI), understanding of 3D scenes for autonomous driving and robotics. An accurate determination of depth within the 3D scene is an important element of these computer vision systems. New emerging applications such as 3D reconstruction, Driver Monitoring Systems (DMS), robotic vision systems for personal robots and advanced HCI modalities require further improvements in short-range depth analysis to better understand and engage with humans.

In this work, we present a method for generating advanced facial models with synthetic data. A method is proposed to generate facial depth information using 3D virtual human and iClone [1] character modelling software. The proposed method can be scaled to produce any number of synthetic facial data by controlling the face animations, scene and camera position.

The main contribution of this research is focused on facial image rendering with the corresponding ground truth depth information. Using the synthetically generated data, we can train CNNs to address the facial depth estimation problem. This approach can enrich the real-world facial datasets required for portrait depth estimation problem.

The rest of the paper is structured as follows: Section II discusses related work and Section III presents the facial models. The application of synthetic facial depth (evaluation) is studied in Section IV. Conclusion and further cautions are discussed in Section V.

## II. RELATED WORK

Facial depth estimation is considered as one of the challenging issues in computer vision, human-computer

interaction and virtual reality. It is used in a wide range of applications which includes controlling 3D avatars, human object detection and human-robot interactions [6-11].

Synthetic human facial data is used frequently to augment real data for pose invariant face recognition. By using the 3D morphable model and Basel face model [13, 14], a pipeline is proposed to create synthetic faces [15]. A synthetic dataset for person identification is studied in [16, 17]. The authors used Blender [2] rendering engine to create different realistic illumination conditions including indoor and outdoor scenes and introduce a novel domain adaptation method that uses the synthetic data. In [13], FaceGen Modeller is used for generating facial ground truth using morphable models. In [19], a large-scale synthetic dataset called (SURREAL) is introduced where the images are rendered from 3D sequences of MoCap data. In [18], synthetic bodies are obtained by utilizing the SMPL body model [18]. This dataset contains more than 6 million frames with ground truth depth, pose and segmentation masks [19].

Very limited work is done on synthetic facial models to explore the field with the available 3D tools and other commercially available software. In this paper, we proposed a method that generates synthetic facial models with many variations in expressions. By controlling the facial animations, camera positions, light positions, body poses, scene illuminations and other scene parameters, the method can be scaled to generate any number of labelled data samples.

### III. FACIAL DEPTH GENERATOR MODEL

Virtual human models are created using the “*Realistic Human 100*” models in iClone [1] software based on the following steps:

#### A. The iClone Character Creation Process

iClone character creator [1] is used to create the initial characters of the virtual human faces. The iClone character creator generates humanoid characters and offers a useful 3D rigging option. The facial animation-ready models can be customized with sculpting and morphs. The template of the “*Realistic Human 100*” models is applied to the base body in the character creator as shown in Fig. 1.



Fig. 1. A sample from the iClone Character creator.

#### B. Adding Facial Expressions to Character Models

The virtual human face models are imported from Character creator to iClone [1]. Further, different expressions are added to the face models to introduce variations such as neutral, angry, happy, sad and scared. Fig. 2, show an example of these expressions.



Fig. 2. A sample rendered images of iClone with different expressions (neutral, angry, happy, sad and scared).

#### C. Exporting Character Animations to Blender

The created virtual human face models are exported from iClone [1] to Blender [2] in FBX format as it provides appropriate rigging. FBX is a popular 3D file format for exchanging the 3D information as used by many 3D tools including Blender [2]. A sample of an iClone facial model with base body loaded in Blender [2] is shown in Fig. 3.

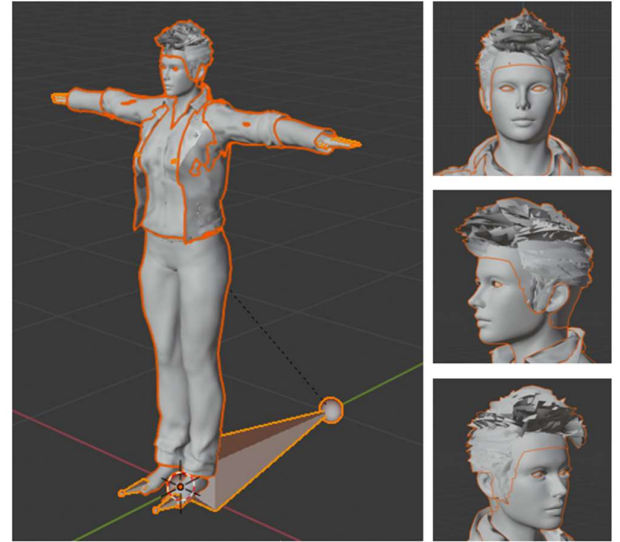


Fig. 3. iClone facial model with base body loaded in Blender.

#### D. Rendering 2D Image Data with Ground Truth Depth

In this work, the following steps are taken to obtain the final output. The cameras and lights are placed in a fixed position and the corresponding distance of the models are changed in the range of 700-1000 mm. The focal length and sensor size are set to 60mm and 36mm respectively. The facial models are rotated in the virtual scenes. Fig. 4 shows a sample

of the camera and light position with respect to the facial models in Blender [2].

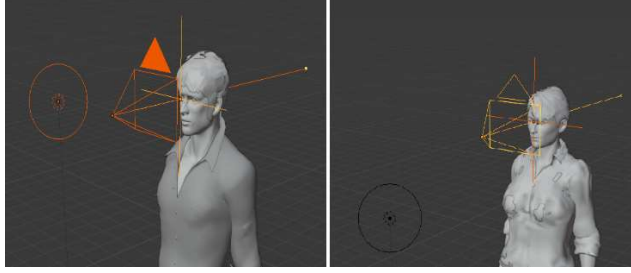


Fig. 4. A sample of the camera and light position with respect to the 3D character.

To generate RGB and depth images of faces in an extensive range of positions, the near and far clip of the camera is set to 0.01 and 5 meters. The facial models are rendered with  $480 \times 640$  resolution and on a static background image. Fig. 5 shows a few rendered images while the camera position is changed with respect to the facial models.



Fig. 5. A facial model with corresponding ground truth depth of a head model from different views.

Fig. 6 illustrates facial models with the corresponding ground truth depth while the camera is positioned at different distances.



Fig. 6. A facial model with ground truth depth captured at the different camera position.

Render passes are set up in Blender [2] to generate the synthetic facial RGB and the corresponding ground truth depth images. To reduce the noise produced during the rendering process, the branched path tracing method is employed. Fig. 7 presents an overview of the noise controlling method in Blender [2].

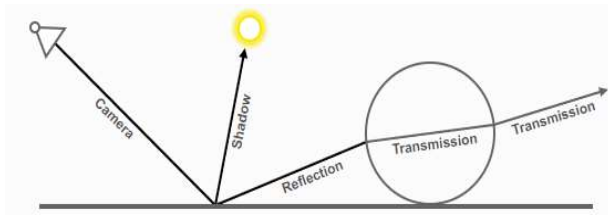


Fig. 7. An overview of the noise control system in Blender.

Afterwards, the images are rendered using Cycles engine and in the perspective view to obtain the RGB images with corresponding facial depth. Fig. 8 demonstrates the workflow of the facial depth generation process, camera and light setting.



Fig. 8. Rendering configuration in Blender. The left row shows the body shape, light and camera setting; the middle row shows the facial RGB and the last row illustrates the corresponding facial depth image.

Fig. 9 shows a few numbers of synthetic male and female models with the corresponding ground truth depth.



Fig. 9. A sample of the synthetic facial images with different expressions and their corresponding depth maps.

#### IV. EVALUATION

In this section, we deliver details about the evaluation of the two-state of the art CNNs on facial depth estimation. The pre-trained monocular depth estimation models DepthDense [19] and MiDas [20] are tested on the rendered synthetic data. Fig. 10, presents a few random synthetic RGB images and the corresponding depth images predicted using DepthDense [19]. Similarly, Fig. 11, shows the synthetic RGB images, predicted depth using MiDas [20] and ground truth images.



Fig. 10. Sample synthetic RGB images predicted depth maps by DepthDense [19] and corresponding ground truth.



TABLE I. RESULTS OF THE DEPTHDENSE, MIDAS MODELS [19, 20] AND SIMPLE CNN MODEL.

No.	Method	Abs Rel	Sq Rel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
1.	DenseDepth[19]	0.8765	0.7783	1.8783	0.2260	0.2723	0.5093	0.6912
2.	MiDas[20]	0.8876	0.9765	1.9876	0.3323	0.3211	0.5432	0.7635
3.	Simple CNN (full image)	0.0412	0.0123	0.0618	0.0177	0.9862	0.9971	0.9989
4.	Simple CNN (only face)	0.0370	0.0092	0.0196	0.0166	0.9961	0.9990	0.9979

<sup>a</sup> Evaluation results of the pre-trained models [19, 20] on the synthetic data.

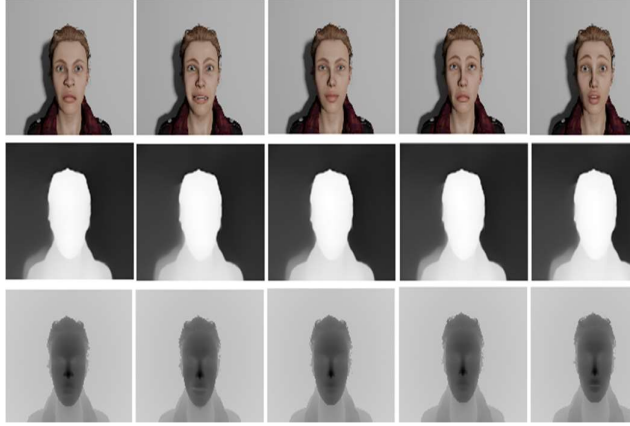


Fig. 11. Sample synthetic RGB images, predicted depth maps by MiDas [20] and corresponding ground truth.

The most common quantitative matrices for evaluating the performance of the pre-trained models including Absolute Relative difference (AbsRel), Root Mean Square Error (RMSE), log Root Mean Square Error (RMSE(log)) and Square Relative error (SqRel) are employed for evaluation purposes. Table 1, demonstrates the evaluation results of the DepthDense and MiDas models [19, 20].

To further evaluate the validity of the synthetic data generated in this paper, we re-trained a few recent CNN-based depth estimation networks [21, 22] on the generated facial data and later fine-tuned the models on real datasets.

A simple autoencoder with skip connection based on U-Net architecture has been trained using the data generated with a plain background as shown in Fig 12.

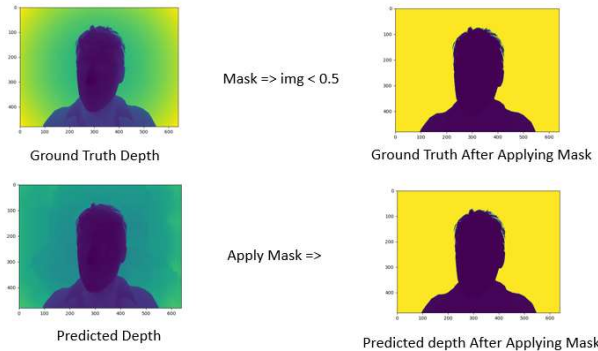


Fig. 12. Ground Truth Depth and Predicted Depth before and applying the mask.

Using the data generated with a plain background as shown in Fig 12, as a monocular depth estimation use case. There are around 40k training and 15k test images and their corresponding ground truth depth. The network has been initialized with random weight and trained with mean square error loss and Adam Optimiser. Further to evaluate the results only on a facial section of the image the depth has been masked within a range of 50 cm from the camera centre and the masked depth has been evaluated with the ground truth depth. Both the results have been shown in Table 1.

Furthermore, we will create additional variations and augmentations in the synthetic facial depth data to grow the final training dataset. It is expected that this will further increase the accuracy of these deep learning-based CNN networks when tested on real data.

## V. CONCLUSION AND FUTURE RESEARCH

In this research paper, we proposed an advanced synthetic facial data generation pipeline. The facial images are generated from 3D virtual human models by rendering different variations of face poses, head poses and lighting conditions. Blender [2] rendering engine is used to generate the output as it allows changing different parameters such as lights position, camera parameters and keyframe values.

The proposed framework has the potential to generate a great number of synthetic facial images. The synthetic 3D models can be used in different 3D environments if scaled properly. This will allow simulating real-world scenarios by controlling the camera position, intrinsic parameters and lighting conditions.

The generated dataset can be used for training and validation of deep learning methods with the focus on natural face modelling, portrait 3D reconstruction and beautification.

In our future work, we will explore the potentials of the deep learning methods on direct facial 3D reconstruction using the synthetically generated data.

## REFERENCES

- [1] 3D Animation Software: iClone: Reallusion. (n.d.). Retrieved from <https://www.reallusion.com/iclone/>.
- [2] Foundation, B. (n.d.). Home of the Blender project - Free and Open 3D Creation Software. Retrieved from <https://www.blender.org/>.
- [3] T. List, J. Bins, J. Vazquez, and R. B. Fisher. "Performance evaluating the evaluator". In ICCCN '05: Proceedings of the 14th International Conference on Computer Communications and Networks, pages 129–136, Washington, DC, USA, 2005. IEEE Computer Society.

- [4] S. R. Musse, R. Rodrigues, M. Paravisi, J. C. S. Jacques. Junior, and C. R. Jung. "Using synthetic ground truth data to evaluate computer vision techniques". In IEEE Workshop on Performance Evaluation of Tracking Systems (in conjunction with ICCV 07), pages 25–32, 2007.
- [5] G. R. Taylor, A. J. Chosak, and P. C. Brewer. Ovvv: "Using virtual worlds to design and evaluate surveillance systems". In Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pages 1–8, 2007.
- [6] S.S. Mukherjee, N.M. Robertson, "Deep head pose: gaze-direction estimation in learning multimodal video", in Proceedings of the TMM, 17, 2015, pp. 2094–2107.
- [7] [7] S. Qi, W. Wang, B. Jia, J. Shen, S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks", in Proceedings of the ECCV, 2018, pp. 401–417.
- [8] Y. Lang, W. Liang, F. Xu, Y. Zhao, L.-F. Yu, "Synthesizing personalized training programs for improving driving habits via virtual reality", in Proceedings of the IEEE Conference on Virtual Reality, 2018.
- [9] C. Li, W. Liang, C. Quigley, Y. Zhao, L.-F. Yu, "Earthquake safety training through virtual drills", in Proceedings of the TVCG, 23(4), 2017, pp. 1275–1284.
- [10] W. Liang, J. Liu, y. Lang, B. Ning, L.-F. Yu, "Functional workspace optimization via learning personal preferences from virtual experiences", in Proceedings of the TVCG, 25(5), 2019, pp. 1836–1845.
- [11] S. Sheikhi, J.-M. Odobez, "Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human-robot interactions", Pattern Recognit. Lett. 66 (2015) 81–90
- [12] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. learning-based, P. van der Smagt, D. Cremers, and T. Brox. "FlowNet: Learning optical flow with convolutional networks". ICCV, 2015.
- [13] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb. "Learning from simulated and unsupervised images through adversarial training". In: CVPR 2017.
- [14] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster and Vetter, T, "Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops pp. 0-0, 2019.
- [15] R. Queiroz, M. Cohen, J. L. Moreira, A. Braun, J. C. J. Júnior & S. R. Musse. "Generating facial ground truth with synthetic faces". In 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images (pp. 25-31). IEEE, 2010.
- [16] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster & T. Vetter. "Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 0-0), 2019.
- [17] Y. Wang, W. Liang, J. Shen, Y. Jia & L. F. Yu. "A deep Coarse-to-Fine network for head pose estimation from synthetic data". Pattern Recognition, 94, 196-206, 2019.
- [18] S. Bak, P. Carr, & J. F. Lalonde. "Domain adaptation through synthesis for unsupervised person re-identification". In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 189-205), 2018.
- [19] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev & C. Schmid. "Learning from synthetic humans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition" (pp. 109-117), 2017.
- [20] I. Alhashim, & P. Wonka. "High-Quality Monocular Depth Estimation via Transfer Learning". 1812.11941, 2018.
- [21] K. Lasinger, R. Ranftl, K. Schindler & V. Koltun. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer". 2019.
- [22] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction". In Proceedings of the IEEE International Conference on Computer Vision, pp. 5684--5693, 2019.
- [23] J. H. Lee, M. K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation". arXiv preprint arXiv:1907.10326, 2019.