



A robust light-weight fused-feature encoder-decoder model for monocular facial depth estimation from single images trained on synthetic data

Title	A robust light-weight fused-feature encoder-decoder model for monocular facial depth estimation from single images trained on synthetic data
Author(s)	Corcoran, Peter;Khan, Faisal;Shariff, Waseem;Farooq, Muhammad Ali;Basak, Shubhajit
Publication Date	2023-04-17
Publisher	IEEE
Repository DOI	https://doi.org/10.1109/ACCESS.2023.3267970

RESEARCH ARTICLE

A Robust Light-Weight Fused-Feature Encoder-Decoder Model for Monocular Facial Depth Estimation From Single Images Trained on Synthetic Data

FAISAL KHAN¹, WASEEM SHARIFF^{1,3}, MUHAMMAD ALI FAROOQ¹,
SHUBHAJIT BASAK², AND PETER CORCORAN¹, (Fellow, IEEE)

¹School of Engineering, National University of Ireland Galway (NUIG), Galway, H91 TK33 Ireland

²School of Computer Science, National University of Ireland Galway (NUIG), Galway, H91 TK33 Ireland

³Xperi Inc., Galway, H91 V0TX Ireland

Corresponding author: Faisal Khan (f.khan4@nuigalway.ie)

This work was supported in part by the College of Science and Engineering, National University of Ireland Galway, Galway, Ireland; and in part by Xperi Galway Block 5, Parkmore East Business Park, Galway.

ABSTRACT Due to the real-time acquisition and reasonable cost of consumer cameras, monocular depth maps have been employed in a variety of visual applications. Regarding ongoing research in depth estimation, they continue to suffer from low accuracy and enormous sensor noise. To improve the prediction of depth maps, this paper proposed a lightweight neural facial depth estimation model based on single image frames. Following a basic encoder-decoder network design, the features are extracted by initializing the encoder with a high-performance pre-trained network and reconstructing high-quality facial depth maps with a simple decoder. The model can employ pixel representations and recover full details in terms of facial features and boundaries by employing a feature fusion module. When tested and evaluated across four public facial depth datasets, the suggested network provides more reliable and state-of-the-art results, with significantly less computational complexity and a reduced number of parameters. The training procedure is primarily based on the use of synthetic human facial images, which provide a consistent ground truth depth map, and the employment of an appropriate loss function leads to higher performance. Numerous experiments have been performed to validate and demonstrate the usefulness of the proposed approach. Finally, the model performs better than existing comparative facial depth networks in terms of generalization ability and robustness across different test datasets, setting a new baseline method for facial depth maps.

INDEX TERMS Facial depth estimation, feature fusion, encoder-decoder architecture, deep learning.

I. INTRODUCTION

Depth estimation is a crucial challenge that is used in a variety of computer vision applications, including 3D vision [1], 3D face recognition [2], and autonomous vehicles [3] due to the low cost of consumer depth cameras and real-time performances. Raw depth maps, on the other hand, continue to face significant acquisition distortion and detailed corruption. An extensive study has lately been conducted to increase

depth accuracy, with the majority of these studies leveraging additional details, such as RGB images or multi-depth maps, for depth map enhancement, while a few employ single depth map enhancement [4], [5], [6], [7], [8]. Although, few studies focus on facial depth maps [9]. The improvement of facial depth estimation is an important research topic for rapid and low-cost 3D facial applications. When compared to ordinary scenery, human faces contain fine structures. Face recognition and other facial depth applications require features that can be used to distinguish one face from another. This makes it more difficult to refine facial depth. With the advancement

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.

of the autonomous industry, it is essential to monitor the driver of a vehicle in order to achieve safety, comfort, and enhanced human-machine interactions [10]. As a proof of concept, the depth estimation in the intelligent vehicle's monitoring system is an advanced way to analyze the driver's behaviour in 3 dimensional instead of 2-dimensional environments. Human facial depth maps are one of the most frequently encountered objects in facial images and are critical for a variety of facial image processing activities. From human facial geometry, the eye separation task in a human facial region is limited to a small range, and thus, using the field of view information from the camera sensors, it is possible to determine the distance between the camera and the subject with reasonable accuracy from a single frame. A neural network can be trained to estimate depth map more accurately by using data that includes face images, which is the main objective of this study. It should be possible for the neural model to understand a considerable measure of the details of human facial structure and properties that can improve the state-of-the-art (SoA) in facial depth map research.

In this paper, the main contribution is to propose a novel neural facial depth estimation network that uses a single image and predicts accurate facial depth maps. As compared to the previous facial depth estimation algorithms, this network is significantly smaller in size and computationally less cost-effective, making it ideal for embedded systems and edge-AI applications. Based on the evaluation of four public facial depth datasets, this lightweight network achieves significantly better results. Furthermore, extensive experiments demonstrate the utility and generalization of the proposed network. The rest of the article is organized as follows. Section II discusses related research that has been conducted in relation to the proposed method. Section III presents and discusses the proposed neural facial depth estimation network for generating facial depth maps. A large and diverse synthetic dataset is used in the training phase, and a series of experimental comparisons, evaluations of the presented approach against the existing SoA approaches and results for facial depth maps are discussed in sections IV and V. In Section VI, the results of this research work are comprehensively discussed. Section VII addresses the challenges, future trends, and improvements, while Section VIII concludes the research.

II. RELATED WORKS

Interpreting spatial relationships within a scene involves estimating depth maps. As a result, such relationships assist in the creation of stronger representations of objects and their surroundings, which can lead to advancements in existing recognition tasks as well as the development of new applications like 3D modelling. With only a single RGB image as input, the purpose of monocular depth estimation is to estimate the depth value of each image pixel or derive a depth map. There has been a lot of effort put into the past to estimate depth using stereo images, as well as progress being made by researchers

in monocular depth estimation due to the advancement in convolutional neural networks [7], [11], [12], [13], [14], [15]. Moreover, with the recent advancements in CNNs and their superior performances, it has been widely used for diversified real-world applications which include speech emotion recognition [42], analysis of non-stationary signals in noisy environments [43], medical image analysis [35], advanced vehicular systems [44] and industrial applications [45]. However, monocular facial depth estimation research has recently gotten attention [9]. Monocular depth estimation employs a single camera to acquire an image or video sequence and requires no more complex equipment or professional techniques. It has a broad range of application requirements due to the availability of only one camera in the majority of application scenarios. As a result, the need for monocular depth estimation has increased in recent years, [15]. Facial depth estimation has many applications and approaches using both conventional and traditional methodologies [8]. Using the feature extraction methods, There are many SoA potential solutions to predict facial depth [16], [17], [18], [19], [20], [21], [22]. Facial feature extraction depth maps can help in the advancement of facial depth tasks. On the other hand, with feature fusion methods, rich internal information of the depth, and compressed reconstructions of integrated features can be generated after dimensionality reduction. There are several approaches that are offered in different tasks: [23], [24], [25].

In recent years, facial depth estimation methods have been proposed for various tasks. Authors in [26], devised a face recognition system in which Fully Convolutional Network (FCN) seeks to recover depth from an RGB image while Convolutional Neural Network (CNN) preserves individual subject separability. In [21] proposed a face depth estimator with conditional generative adversarial networks (GAN). They created a GAN-based approach for estimating depth maps from single-face images. This method also concluded that the conditional Wasserstein GAN structure is the most reliable technique using GAN-based networks. Authors in [27] used an unsupervised approach to estimate depth with 3D face rotation and replacement by implying the depth of an input image's facial key points. In [28] proposed a GAN-based technique to produce robust facial depth estimation. Further [9] proposed a GAN-based technique via segmentation and mask-guided attention network for face depth estimation. Recent research has also revealed that, in addition to colour and deformation, the depth of Ground Truth (GT) of a face can be used to discriminate between real and synthetic faces. It is a strategy worth researching to increase the label information by utilizing estimated depth image labels instead of coding labels. The authors in [30], suggested an auxiliary supervised technique that uses estimated face depth information to expand label information.

In addition, there has been significant research towards generating 3D synthetic facial depth estimation methodologies. In [31], authors provided realistic 2D facial depth models obtained from a 3D synthetic dataset. The authors also

suggested a benchmark dataset, as well as a CNN-based architecture for predicting depth from a 2D image in [32]. The authors in [8] offered a comprehensive review of monocular facial depth estimation, including types of approaches that have been and can be used in past, current, and future research.

III. LIGHTWEIGHT ENCODER-DECODER BASED FACIAL DEPTH ESTIMATION MODEL

Numerous consumer applications, such as robots, augmented reality, and automated driver monitoring systems, can benefit from neural facial depth estimation networks constructed from single image frames. Conventional approaches utilize fully connected layers, which complicates the models and necessitates additional memory, making them unsuitable for deployment on consumer devices and they suffer from issues such as information loss that leads to holes in depth-images. On the other hand, many Deep Learning (DL) techniques have recently been presented, and they have shown considerable progress in solving the fundamental ill-posed problem of depth estimation. This article describes the procedure for constructing depth maps from a single-frame face image that makes use of the input RGB face image and the corresponding GT depth utilizing neural networks.

Keeping a simple model architecture in mind that can be used for consumer devices for real-life applications, the model applied in this research work automates the collection of optimal parameters and a less number of parameters size thus reducing model complexity during the training procedure. The proposed model is more computationally efficient than the current SoA facial depth maps models and shows performance equal to, or better than SoA when tested across 4 public depth datasets. The performance of the proposed CNN model is evaluated with SoA networks, and different encoders including EfficientNetB0, ResNet-101, and DenseNet-169 are compared.

A. NETWORK ARCHITECTURE

This section describes the proposed neural facial depth network for the mechanism of single-image facial depth maps, as well as the suggested loss function for optimizing the procedure over the training data. The framework's general architecture is demonstrated in Fig. 1. To obtain high-quality facial depth maps, researchers usually create deeper networks with additional parameters and constraints, which need additional computation complexity and hence do not match the real-time requirements of real-time applications. As a result, the authors sought to develop a lightweight neural facial depth model capable of real-time facial depth prediction while maintaining prediction accuracy equal to or better than current SoA networks.

1) ENCODER MODEL

The proposed decoder for reconstructing facial depth residuals is coupled to the network's pre-trained encoder ResNet18 [33] and the main feature of the network have been

described in Fig. 1 and table 1. In the encoder process, the model consists of 22 layers including eight parts: convolutional layers 1-5, a global average pooling (AP) layer, and a fully connected (FC) layer. The initial features are corrected in the channel dimension to increase the model's intensity of learning features, enabling the model to automatically pick up on the key characteristics of various channels. The global average pooling layer is then used in place of the fully-connected layers to decrease model parameters, speed up model convergence and enhance the accuracy of the model.

TABLE 1. The Encoder model's detailed structure is used in the proposed method.

Layers	Output Size	Layer Parameters
C1	112x112x64	7x7,64, stride 2
C2	56x56x64	3x3 maxpool, stride 2 3x3,64 x2 3x3,64
C3	28x28x128	3x3,128 x2 3x3,128
C4	14x14x256	3x3,256 x2 3x3,256
C5	7x7x512	3x3,512 x2 3x3,512
AP	1x1x512	7x7 AP
FC	2	512x2 FC
softmax	2	FC

2) DECODER MODEL

The presented model's encoder takes the input image to block features of various sizes. A lightweight and efficient decoder is utilized to recover the bottleneck features in order to extract the estimated facial depth map [6]. The better performance is demonstrated experimentally to be due to the training process. Additionally, the model achieves higher performance by utilizing much fewer convolutional and bilinear upsampling layers in the decoder. To begin, convolution is employed to lower the channel dimension of the bottleneck feature, hence avoiding the complexity of the algorithm. Following that, a series of bilinear upsampling layers are utilized to enhance the size of the features. Lastly, two convolution layers and a sigmoid function are used to the output to estimate the facial depth map. Additionally, the depth map is scaled by the value of the maximum depth to give the depth in meters. A skip connection is introduced to the proposed fusion module in order to make better use of the precise details of the local structures.

B. LOSS FUNCTION

The objective of the facial depth estimation problem is to design a function that accurately predicts the depth of an input image. (L_{silog}) seems to be the most frequently used and the best choice loss function in the training process because it is more useful for reducing errors in facial depth estimation.

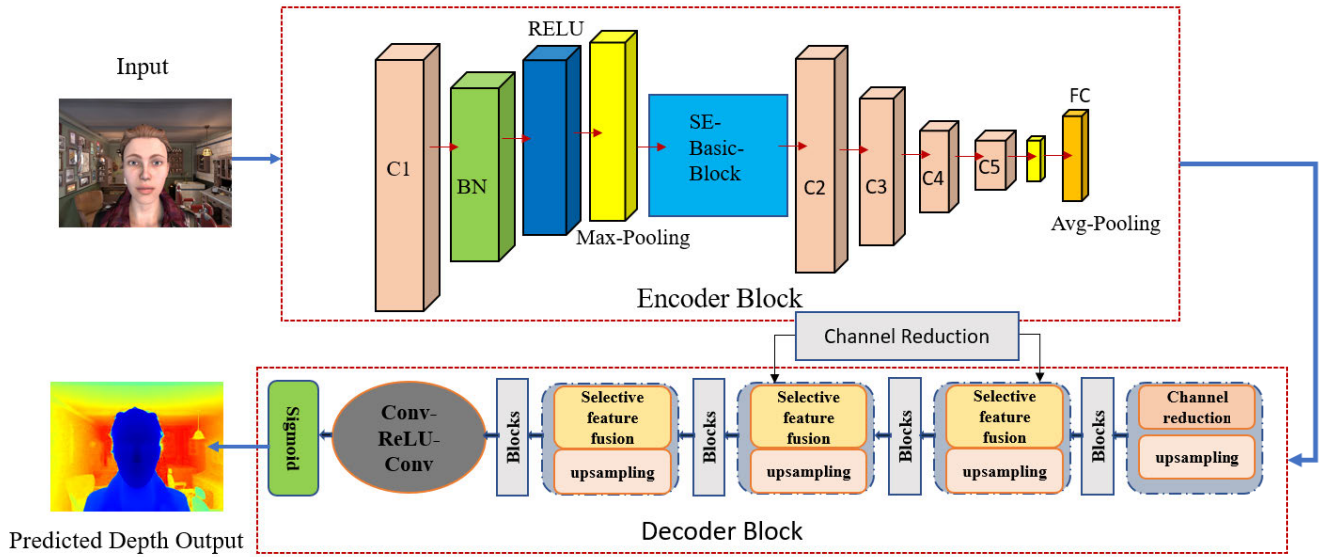


FIGURE 1. The proposed approach for monocular facial depth estimation’s architectural shape. The encoder has the Resnet18 network, and the proposed decoder architecture’s primary components along with channel reductions, skip connections and feature fusion modules.

The network’s learnable parameters are optimized focusing on the loss function, which implements correctly scaling the loss function’s range to enhance convergence and training output results while attempting to put a stronger focus on *lamda*-based error variance reduction, resulting in a Silog loss function. Reference [14](Lee, Han, Ko and Suh, 2019b) (L_{silog}) is defined:

$$L(y_i, y_i^*) = \frac{1}{n} \sum_i^n (\log(y_i) - \log(y_i^*))^2 - \frac{\hat{\lambda}}{n^2} \left(\sum_i^n \log(y_i) - \log(y_i^*) \right)^2 \quad (1)$$

where $\hat{\lambda}$ is the balancing factor, and n is the pixel count. Through a rewrite of the equation. 1:

$$L(y_i, y_i^*) = \frac{1}{n} \sum_i^n (\log(y_i) - \log(y_i^*))^2 - \left(\frac{\hat{\lambda}}{n} \sum_i^n (\log(y_i) - \log(y_i^*)) \right)^2 + (1 - \hat{\lambda}) \left(\frac{1}{n} \sum_i^n (\log(y_i) - \log(y_i^*)) \right)^2 \quad (2)$$

It’s a sum of the variance and a balanced square average of the error in log space. As a result, founding a larger $\hat{\lambda}$ imposes a greater focus on limiting error variance, Also, it is found that adjusting the loss function’s range properly increases convergence and the overall training result. In log space, the combined Silog loss is defined as:

$$L_{silog}(y_i, y_i^*) = \alpha \sqrt{L_{(y_i, y_i^*)}} \quad (3)$$

IV. EXPERIMENTS

The experimental results are discussed and summarized to demonstrate the effectiveness of the proposed approach in comparison to SoA methods. The proposed model is trained on a synthetic facial depth dataset and then compared to four real datasets. Numerous comparisons have been conducted, as well as evaluations of its accuracy and computational footprint.

The studies show that a network trained on a wide and diverse set of images, along with a decent training technique, produces SoA performance in many situations, particularly for faces. The zero-shot cross-dataset transfer technique is used to show the method’s effectiveness.

A. IMPLEMENTATION DETAILS

The model for estimating the facial depth is trained with the PyTorch DL framework. For training and testing, the data was divided into 0.8 and 0.2 ratios, and the model was evaluated against four publicly available datasets. We employ the one-cycle learning rate technique with an Adam optimizer in all of the experiments. The learning rate increased by 0.9 during the first half of the total iterations from 3e-5 to 1e-4 following a poly LR scheduling and then falls by a factor of 0.9 from 1e-4 to 3e-5 in the second half. On a workstation equipped with NVIDIA 2080ti GPUs, the total number of epochs is set to 50 with a batch size of 16. There are around 12.06 million trainable parameters in the proposed model.

The Root Mean Square Error (RMSE), the log Root Mean Square Error (RMSE (log)), the Absolute Relative difference (AbsRel), the Square Relative error (SqRel), and the Accuracies are used to perform the evaluations (Equation (4-10)). With a 50% probability, the following procedures are utilized for data augmentation: horizontal flips, random

brightness(0.2), contrast(0.2), gamma(20), hue(20), saturation(30), and value(20). We use $p = 0.75$ for vertical CutDepth with a probability of 25%.

Fig. 2 depicts the whole experimental implementation details including training, evaluation, and testing of the proposed model using a synthetic facial depth dataset. First, the model is trained with a synthetic facial depth dataset and then evaluated and tested with four real depth datasets (mentioned in section IV-C) against SoA DL methods. The proposed model uses a single frame RGB image and corresponding GT depth image as training data for the convolution layer to extract features. CNN uses a weight-sharing method that significantly reduces the number of parameters, greatly enhancing the model's performance.

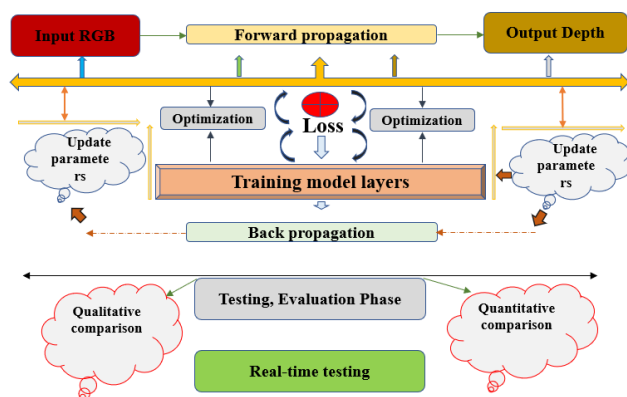


FIGURE 2. The implementation details of the suggested neural network training, evaluation and testing procedure.

B. TRAINING DATASET

The proposed LEDDEPTH model is trained on the synthetic human facial depth dataset and evaluated with four other test datasets for rigorous comparison with other SoA models which includes BTS [14], Densdepth [34], UNet-simple [32], ResNet-101 [36], EfficientNet-B0 [37], MiDaS [5]. Further details of the training dataset are presented in sub-section IV-B1.

1) SYNTHETIC HUMAN FACIAL DEPTH DATASET [36]

There are a considerable number of high-quality 3D face models in the Synthetic Human facial depth datasets as well as 2D RGB and pixel-accurate ground truth depth images. Character Creator is accustomed to using 100 real-world head models to create a series of virtual human avatars. The models' textures and topologies are adjusted to increase the number of possible samples. After loading the models into iClone, 5 distinct facial expressions are incorporated into the data. Importing the FBX files of the iClone models and their associated mesh, textures, and animation keyframes into Blender is the final step in the process.

All Blender models have been rotated in order to get the proper head position. Thereafter, the FBX models are imported into Blender and adjusted to the reference frame.

Lights and cameras are used in the environment to mimic the real-world environment, and their attributes are then altered accordingly. The camera lens's near and far clips have been set to a distance of 0.01 meters and a maximum of 5 meters. 60 degrees of FOV is achieved by adjusting both the sensor's resolution and the sensor's field of view (FOV). The final effect is attained by configuring the render layer's RGB and Z-pass outputs in the compositor. In posture mode, the joints of the head and shoulders are detected, the head mesh pivots these bones, and frames are saved to carry out the rotational movement.

Finally, all frames are rendered in order to produce the RGB and depth images needed for final rendering. With the help of the Python code available in the Blender application, the head position (yaw, pitch, and roll) has been created. A 640×480 pixel RGB image is created and saved in jpg format for each frame. while the depth data is saved in (.exr format). A text file (.txt) containing each frame's head positions is also stored. The Cycle Rendering Engine, integrated with Blender, renders each 2D image in about 26.3 seconds on average. Using Cycle Rendering Engines is used to track the progress of the rendered scenes. There are around 3,500k frames in the entire collection, and each model receives about 3.5k 2D images. The following link has detailed information about the dataset. (<https://dx.doi.org/10.21227/ath9-br59>)

C. TEST DATASETS

There are numerous datasets available for estimating facial depth, each with a unique type and depth range. Four datasets are chosen for the diversity and quality of their source data for facial depth map predictions. Those include the following: Pandora [38], Eurecom Kinect Face [39], Biwi Kinect Head Pose [39] and Synthetic Human Facial Depth [36] test dataset for testing and evaluations purposes.

1) PANDORA

The Pandora dataset is utilized for a variety of purposes, including estimating head pose, head centre localisation, depth estimation, and shoulders pose estimation. It includes 250K full-resolution RGB images and their corresponding depth images.

2) EURECOM KINECT FACE

The dataset contains multi-model facial images of 52 individuals, 38 of who are male and 14 of whom are female, collected with the Kinect sensor. It includes nine distinct states of facial expression, occlusion, and illumination, including grin, eye obstruction, mouth, light and sheet, moderate, open mouth, and left-right profiling.

3) BIWI KINECT HEAD POSE

Contains 15k images of 20 subjects taken with the Kinect sensor as the subjects' heads were freely moved around

TABLE 2. Comparison of various depth maps methods with the proposed method LEDDEPTH, BTS [14], Densedept [34], UNet-simple [32], ResNet-101 [36], EfficientNet-B0 [37], MiDaS [5], DPT [15], LapDepth-Face [8], FaceDepth [36] on synthetic human facial depth dataset [36].

No.	Methods	AbsRel	SqRel	RMSE	RMSElog	$\delta_1 < 1.25$	$\delta_2 < 1.25^2$	$\delta_3 < 1.25^3$
1.	DenseDepth-161	0.0296	0.0096	0.0373	0.0129	0.9890	0.9920	0.9981
2.	ResNet-101	0.0123	0.0210	0.0306	0.0089	0.9938	0.9960	0.9980
3.	BTS	0.0165	0.0092	0.0206	0.0102	0.9830	0.9943	0.9956
4.	EfficientNet-B0	0.0145	0.0280	0.0360	0.0154	0.9912	0.9934	0.9978
5.	UNet-simple	0.0103	0.0207	0.0281	0.0089	0.9960	0.9956	0.9987
6.	MiDaS	0.0146	0.0204	0.0356	0.0323	0.9665	0.9902	0.9983
7.	DPT	0.0156	0.0106	0.0394	0.0184	0.9567	0.9646	0.9943
8.	LapDepth-Face	0.0145	0.0041	0.0204	0.3614	0.9545	0.9857	0.9958
9.	FaceDepth	0.0176	0.0030	0.0205	0.1252	0.9642	0.9849	0.9951
10.	LEDDEPTH	0.0113	0.0025	0.0203	0.1172	0.9888	0.9961	0.9967

across each side. Each frame contains RGB and depth images, as well as the head's 3D position and rotation angles.

D. EVALUATION METRICS

To interpret the data, a widely known assessment procedure with several evaluation indicators is being used: The root mean square error (RMSE), the log root mean square error (RMSE (log)), the absolute relative difference (AbsRel), the square relative error (SqRel), the accuracies, the normalized root mean square error (NRMSE), and the R-squared. All of those are as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \|y_i - y_i^*\|^2} \quad (4)$$

$$RMSELog = \frac{1}{n} \sum_{i=1}^n \|\log(y_i) - \log(y_i^*)\|^2 \quad (5)$$

$$AbsRel = \frac{1}{n} \sum_{i=1}^n \frac{\|y_i - y_i^*\|}{y_i^*} \quad (6)$$

$$SqRel = \frac{1}{n} \sum_{i=1}^n \frac{\|y_i - y_i^*\|^2}{y_i^*} \quad (7)$$

$$Accuracies = \% \text{ of } y_i \max\left(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}\right) = \delta < thr \quad (8)$$

$$NRMSE = \frac{RMSE - RMSE_{min}}{RMSE_{max} - RMSE_{min}} \quad (9)$$

$$R^2 = 1 - \frac{\sum_{m=1}^n (y_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - y_i^*)^2} \quad (10)$$

where y_i^* is the GT, \bar{y}_i^* is the mean of the GT and y_i is the predicted depth of the pixel i , n represents the overall number of pixels, while thr denotes the accuracy threshold.

V. RESULTS AND COMPARISONS TO PRIOR WORK

The results of the proposed approach are shown in Fig. 3 and Table 2. The performance of the proposed facial depth estimation model is evaluated with the SoA methods BTS [14]; MiDaS [5]; DPT [15]; LapDepth-Face [8] on the synthetic human facial dataset [36]. The network achieves SoA performances in the evaluation metrics SqRel, RMSE and δ_2 .

For depth map estimation RMSE is considered the most focal metric for loss estimation thus measuring the performance evaluation of the depth architectures. As can be observed from Table 2, the proposed architecture outperforms other SoA Depth models having the lowest RMSE value.

In the evaluated matrices in Table 2, it can also be observed that the Unet-simple model performs better or is comparable to the suggested model in AbsRel, RMSElog, and δ_1 . The main reason for these results is that the model was trained across the entire image first before being applied to the Facial crop (FC) for evaluating errors in the face region. In other words, the depth has been masked within a 50-centimetre range from the camera so that the results can only be evaluated on the facial region of the images.

The results, as shown in Fig. 4, display high-level detail and constancy, showing that the suggested method performs better at estimating facial depth maps. Note: due to the fact that the MiDaS network was built to predict inverse depth, the predicted images differ from those of other SoA. Fig. 7 demonstrate the proposed model's qualitative results on real data and synthetic data compared to SoA techniques. The model outperforms the cutting-edge techniques and sets a new SoA for facial depth estimation. According to the comparison study Table 2 and Fig. 4, the proposed LedDepth method performed best in terms of accuracy and depth range when compared to other SoA approaches. On a synthetic human facial dataset, the network achieved 0.0203 RMSE and 0.9986 threshold accuracy. To the SoA approaches, the suggested lightweight network structure has less parameters and complexity and can be seen from Table 3 and Fig. 5, which provides a full comparative analysis in terms of the number of parameters and computational complexity.

A. QUALITATIVE RESULT

In this subsection, the authors compare qualitative results from the proposed model to SoA approaches. A comprehensive analysis of the proposed method to the four best-performing methods is shown in Fig. 4 and Fig. 6. The suggested model results show better information and consistency, as shown in Fig. 6, proving that the network works better at facial depth estimation.

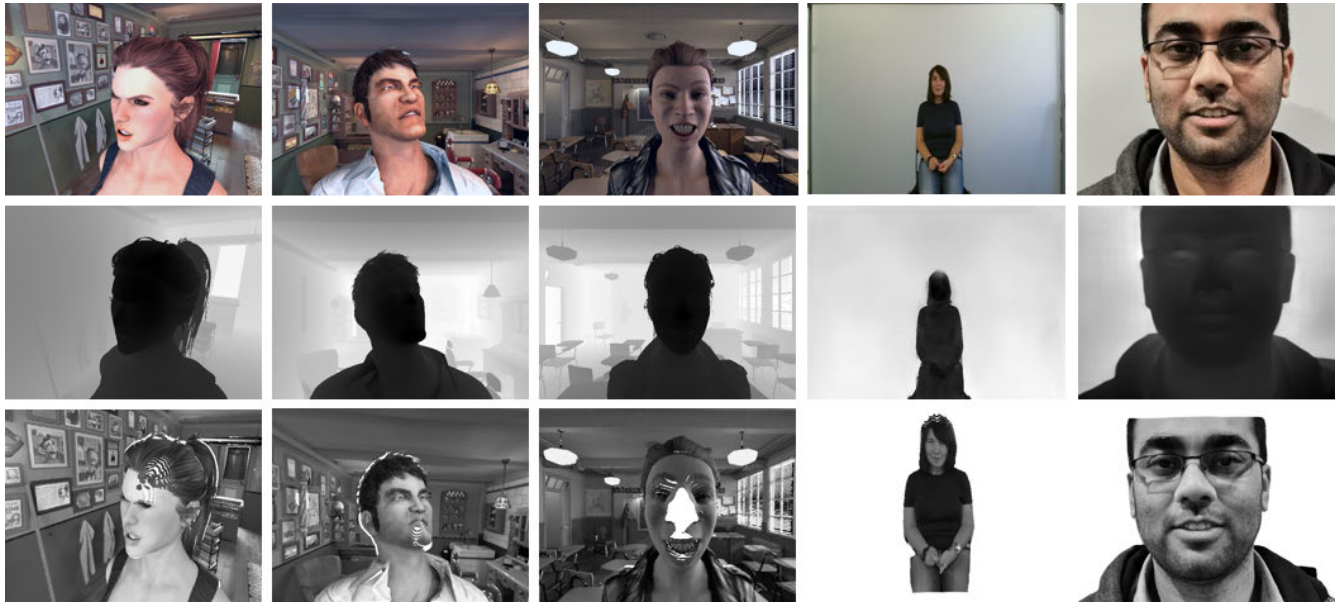


FIGURE 3. The suggested method was evaluated qualitatively using a sample of the synthetic human facial data that was not utilized for training or validation. The first row consists of input RGB images, the second row consists of corresponding predicted depth images, and lastly their rendered point clouds from a novel viewpoint. Point clouds rendered via Open3D [41].

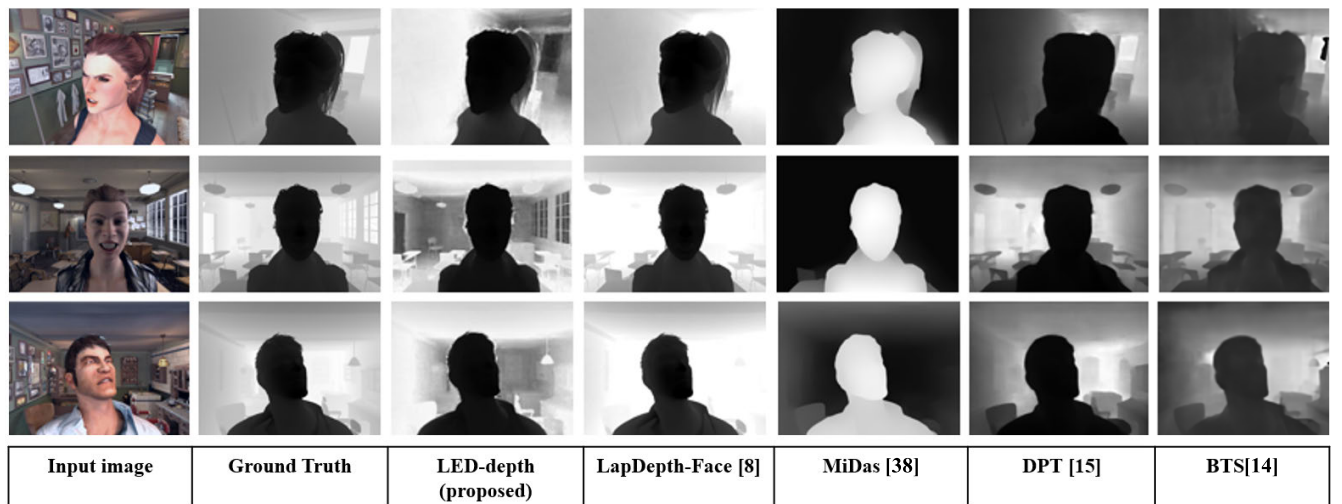


FIGURE 4. Qualitative results of facial monocular depth estimation algorithms on the synthetic human facial dataset.

The model outperformed SoA quantitatively and qualitatively in testing using four datasets and formed a new SoA for facial depth maps. Table 2, Table 3, and Fig. 6 illustrate some of the results.

According to the analyses, the presented scheme significantly outperforms other SoA methods on the basis of consistency and depth range. On the synthetic human facial data, the neural framework obtained a SqRel of 0.0025, RMSE of 0.0203 and a threshold accuracy of 0.9961 as can be seen in Table 1 (row 10).

Additionally, as demonstrated in Table 3 (row 10), the suggested method has a much smaller memory footprint and higher computational efficiency when compared to previous

SoA methods. At 25.32 G-MACs per image, this technique enables real-time prediction of single image face depth. Although the LedDepth model has fewer parameters than other SoA, the design principle and simple encoder-decoder stages make it computationally less expensive and can be used for consumer devices.

The following Table 3 summarizes the characteristics of the models for predicting facial depth maps of a single image frame that have been studied and compared (ED: Encoder-Decoder; F: Trained on the synthetic human facial dataset). According to the test results, DPT [15]; MiDaS [5]; LapDepth-Face [8]; BTS [14] and FaceDepth [36] techniques can build high-resolution facial depth maps with comparable

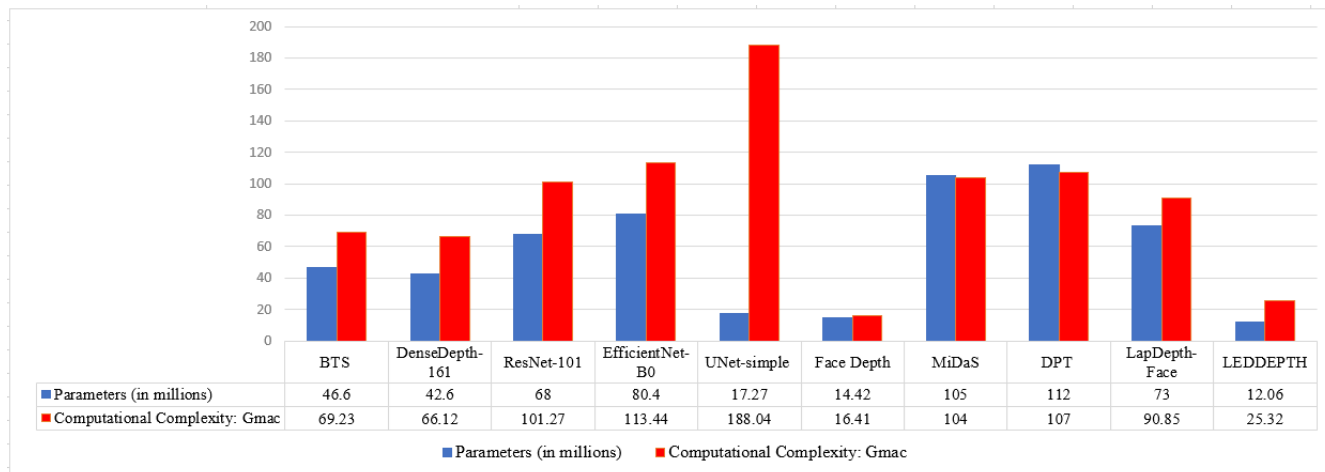


FIGURE 5. The comparison of parameters and their cumulative sum. The proposed LEDDEPTH model contains much less parameters, as shown by the cumulative percentage.

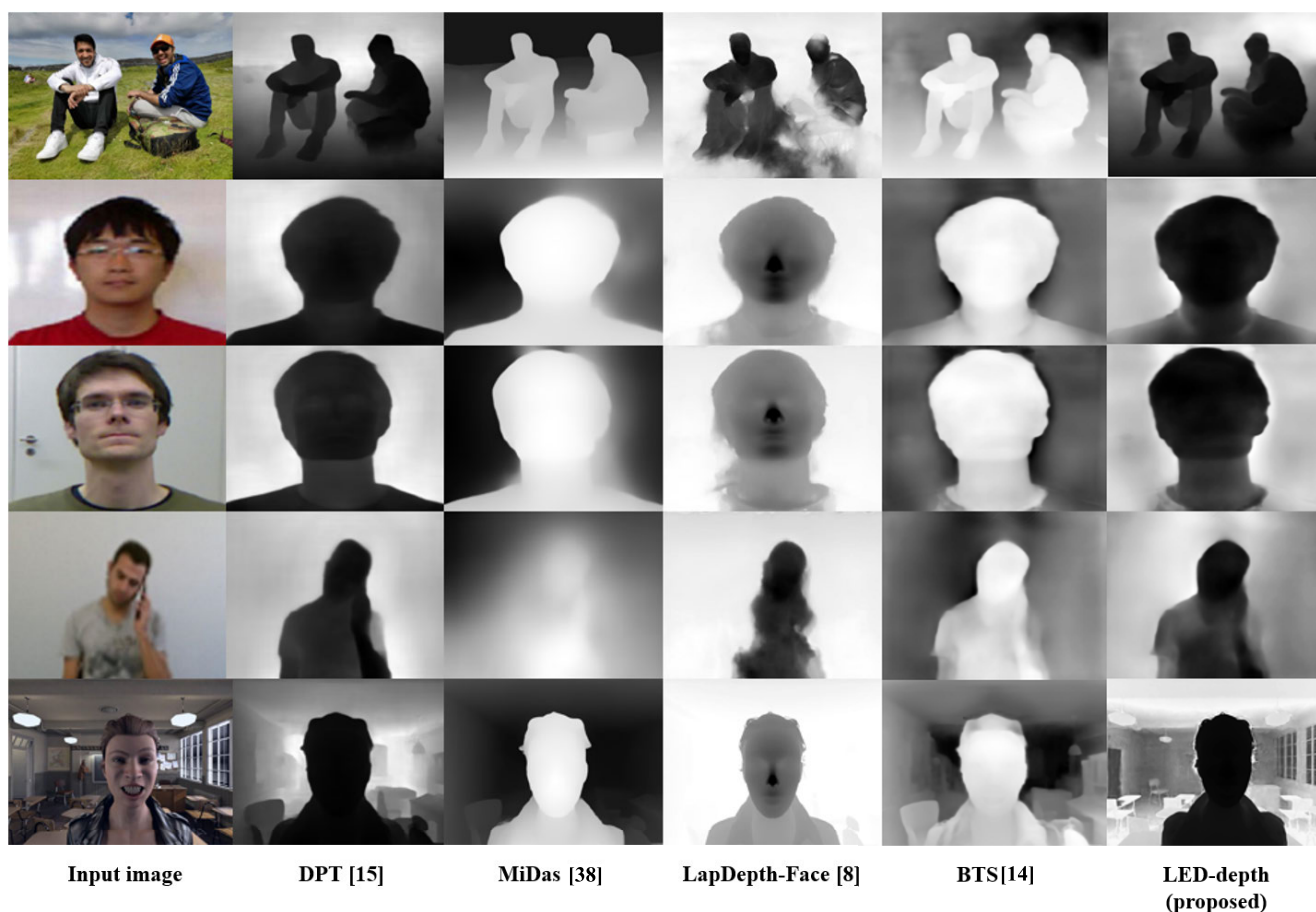


FIGURE 6. A qualitative analysis of our technique in relation to the four SoA methods applied to different datasets (From below:- Synthetic human facial dataset [36]; Pandora dataset [38]; Eurecom Kinect Face dataset [39]; Biwi Kinect Head Pose dataset [40] and an image taken from iPhone 13 pro.

accuracy but are computationally expensive and require a large amount of memory. On the other hand, LedDepth significantly reduced computation time and memory footprints, making it suitable for both high-quality and low-cost single-image facial depth estimation (Table 3 and Fig. 6).

VI. DISCUSSION

This research proposes a neural model for facial depth estimation and compares its performance to that of current SoA algorithms. Compared to other SoA techniques, the framework proposed has a significantly smaller network size,

TABLE 3. Properties of the studied methods with the proposed method LEDDEPTH, (ED: Encoder-Decoder; F: Trained on the synthetic human facial dataset); LR/E: Learning Rate/Epochs; CC: Computational Complexity.

Method	Input	Type	Optimizer	Parameters	Output	LR/E	CC (GMac)
BTS [14]	640×480F	ED	Adam	46.60M	640×480F	0.0001/50	69.23
DenseDepth-169 [34]	640×480F	ED	Adam	42.60M	320×240F	0.0001/20	66.12
ResNet-101 [36]	640×480F	ED	Adam	68.00M	640×480F	0.0001/25	101.27
EfficientNet-B0 [37]	640×480F	ED	Adam	80.40M	640×480F	0.00001/20	113.44
UNet-simple [32]	640×480F	UNet	Adam	17.27M	640×480F	0.001/20	188.04
FaceDepth [36]	640×480F	ED	Adam	14.42M	320×240F	0.0001/50	16.41
MiDaS [5]	384×384F	CNN	Adam	105.00M	384×384F	0.0001/60	104.00
DPT [15]	384×384F	Transformer	Adam	112.00M	384×384F	0.00001/60	107.00
LapDepth-Face [8]	512×416F	ED	Adam	73.00M	512×416F	0.00001/50	90.85
LEDDEPTH (Proposed)	640×480F	ED	Adam	12.06M	640×480F	0.0001/50	25.32

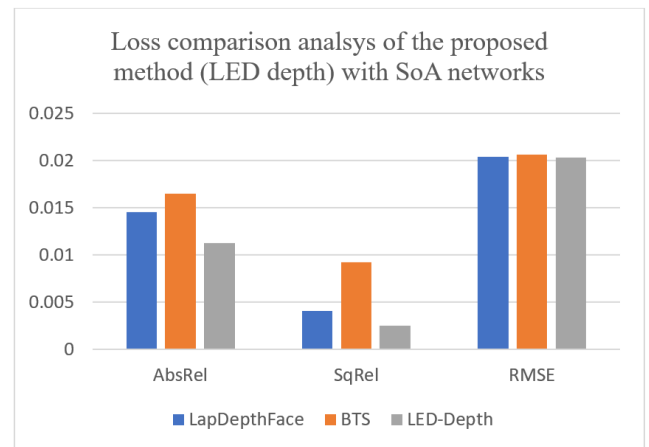
a smaller number of parameters, and equal or better computing complexity (only less than UNet-simple [32]). It can be noted that when compared to the proposed LEDDEPTH approach, the FaceDepth method in Table 3 is superior in terms of computational complexity, however, the qualitative results and evaluation metrics are superior to the FaceDepth method in Table 1. In comparison to the existing FaceDepth, the suggested LEDDEPT performed best in terms of accuracy and depth range and can improve its performance in different testing scenes.

The usefulness of the performance is related to the neural model training strategy, which chooses an appropriately optimal loss function by utilizing a synthetic human facial dataset with pixel-accurate ground truth depth information.

As seen in Table 2 and Table 3, the suggested model performs well on the majority of evaluation criteria (Table 1 row 10), which the authors explain to the proposed design and improved depth-specific training methods. Additionally, the suggested neural model outperforms recently published SoA algorithms with fewer parameters (Table 3 blue row 10). This indicates that the integration of the encoder and the suggested simple decoder clearly contributes significantly to the fast obtaining of accurate facial depth maps. Fig. 6 depicts the visible results. As illustrated in the figure, the model accurately estimates facial depth values for the sample images and is more robust to changing illumination circumstances than other methods. In terms of generalization, DPT and MiDaS performed well in some test images for long-range recognition, however, the proposed LEDDEPTH technique performed better for short-range attribution, particularly for facial regions.

We perform all tests and evaluations of the SoA on a set of datasets that were never seen during training for both real imaged and synthetic datasets and the results are evaluated using Equations (4-10).

Fig. 7 shows a visual representation of the three different loss function comparisons of the proposed model with two SoA depth networks (BTS and LapDepthFace) It is obvious that the suggested method achieves good performance by minimizing errors over a large number of test datasets when compared to other SoA algorithms. By selecting an appropriate loss function and a pixel-accurate synthetic facial depth

**FIGURE 7.** The three different evaluations errors metrics: AbsRel, SqRel and RMSE comparison between the proposed network with two SoA networks.

dataset, the algorithm is able to decrease error while having a small number of parameters and equal or less computation complexity.

Furthermore, The proposed model is converted to ONNX and it can be used for deployment in embedded systems and in Edge-AI applications. ONNX is a freely available format for encoding deep neural networks. With ONNX, Application developers can more quickly integrate models between SoA packages and determine the ideal mix for their needs. A community of contributors contributes to the development and support of ONNX. Lastly, the release of the code utilized in this study and the publicly available training dataset, as well as the corresponding ONNX transformations, will aid future research in fields like as 3D facial reconstruction, perception, and characterization. https://github.com/khan9048/Facial_Depth_Maps_from_Single_Images

VII. CHALLENGES AND TRENDS

Monocular depth estimation based on DL has been widely researched and advanced during the previous decades. Nevertheless, much more work is required to overcome the limitations, particularly in the area of facial depth estimation.

To enhance the accuracy of depth maps, the majority of studies have concentrated on the layers of neural models, which increases the capacity of the space model and memory consumption. In multi-task neural depth methods for monocular facial depth maps usually use numerous sub-networks to execute distinct sub-tasks, which also increases computations and memory requirements. Typically, most of the monocular facial depth estimation networks are encoder-decoders with complex structures. After numerous levels of information computation, the depth characteristics are significantly degraded, leading to decreased estimated depth maps that do not fulfil the practical requirements of the application.

This section covers the major issues and discusses potential directions for monocular facial depth estimation research that can help the researchers in further developments.

A. HIGH-RESOLUTION DEPTH MAP OUTPUT

Facial depth estimation is a critical phase in the evolution of real-world applications such as augmented reality (AR) and virtual reality (VR), and it imposes a great deal of importance on the depth maps accuracies. Nonetheless, the quality of the anticipated facial depth is often limited in most contemporary algorithms in order to maximize computational effectiveness. At the current, research studies are enhancing the super-resolution of depth images using colour image super-resolution frameworks. However, how to properly produce a high-resolution facial depth map remains an open question.

B. REAL-TIME PERFORMANCE

The fundamental module of SLAM is image depth maps, which are tightly coupled with industrial applications such as autonomous driving. As a result, practical applications required pixel-accurate depth map performance. However, in order to achieve high-quality depth maps, researchers frequently design deeper networks with more parameters and requirements, which requires more computation time and thus does not meet the real-time requirements of real-time applications. Thus, a future research area will be to determine how to use lightweight neural depth models for real-time depth prediction while maintaining prediction accuracy.

C. INTEGRATION AND OPTIMIZATION OF THE NETWORK FRAMEWORK

While it is possible to combine or build a network that can learn both facial depth and segmentation in DL facial depth estimation research, this remains a distinct research field. To learn several tasks, such as face depth maps or segmentation or depth features or optical flow prediction and visual odometry simultaneously, sub-models are typically used in an unsupervised manner. These models, however, are not effectively integrated, which results in a high number of parameters, which increases the memory needs and computational complexity of the system. The neural model needs to be better integrated, and this is a research topic worth pursuing in the future.

With a DL model, we may acquire several features at once, such as semantics, optical flow features as well as depth information. Different aspects are obtained and matched simultaneously during the encoding stage; they are decoded independently to meet the requirements of the applications during the decoding step.

D. DYNAMIC OBJECTS AND OCCLUSION PROBLEMS

In order to create realistic scenes, developers must consider a range of aspects, such as a large group of moving parts, occlusions, shifting lighting, and varying weather. Most existing facial depth estimation algorithms, on the other hand, simply take into account ideal circumstances. Researchers have made progress in recent years in dealing with moving objects and occlusion environments, but the challenge of accurately estimating the facial depth of complicated environments to satisfy real-world applications remains a major challenge.

E. DATASETS CONSTRUCTION

The consistency and generalization of a learning algorithm are heavily influenced by the quality of the datasets used to train it. Facial depth maps can be improved if more data, with greater quality, and more scene types are available. These available datasets for facial depth maps are limited, and the production of a new dataset is time-consuming and costly. Currently, some researchers are using computers to make a larger number of images for depth maps, but the quality is unstable. In the future, researchers will be looking at how to build a dataset for a monocular face depth map that is suitable for DL.

For instance, synthetic human facial data generation can give better ground truth depth information than can be collected in practice, so high-quality training data can be utilized to produce better single image depth algorithms. Adding real-data samples, enhancing the hyperrealism of the synthetic datasets, and including a larger range of face characteristics, emotions, and scene illumination could allow for further progress beyond SoA.

VIII. CONCLUSION

The main contribution of this paper is a new lightweight neural facial depth estimation network based on a single image depth map. While this network is compatible with previous SoA facial depth estimation techniques, it is substantially smaller in size and computation cost, making it suitable for embedded devices and edge-AI applications. When evaluated over four publicly available datasets, this model outperforms SoA on most of the primary measures including RMSE, SqRel and δ_2 . Furthermore, comprehensive experiments show the proposed network's robustness and generalization capability.

A crucial aspect of this research is that training neural facial depth networks on synthetic human facial data produces higher-quality depth maps than is possible through the available realistic datasets. Using lightweight neural single-image depth predictions, high-quality training data may be used

to generate accurate facial depth maps. More optimizations beyond SoA should be possible through the incorporation of large and diverse facial depth datasets. Obviously, synthetic facial data will lack the richness of real facial and skin features coming from the real dataset. However, considering the numerous benefits of training a neural depth model with synthetic data, a critical research question is whether it is possible to accomplish comparable results that are answered to SoA facial depth estimation models trained on real-world data. It is possible that future research can include exploration towards high-resolution facial depth maps, system integration and optimization, high-resolution facial depth map efficiency, data augmentation methods and analyses with a wider range of sample datasets. It would be interesting to investigate a combined multi-tasks network to specifically address downstream applications including image classification, depth maps prediction and semantic segmentation. Another potential future study dimension is multi-frame facial depth, which leverages a succession of image frames and may be paired with some motion estimation or disparity information.

REFERENCES

- [1] R. Xiong, S. Zhang, Z. Gan, Z. Qi, M. Liu, X. Xu, Q. Wang, J. Zhang, F. Li, and X. Chen, "A novel 3D-vision-based collaborative robot as a scope holding system for port surgery: A technical feasibility study," *Neurosurgical Focus*, vol. 52, no. 1, p. E13, Jan. 2022.
- [2] M. Li, B. Huang, and G. Tian, "A comprehensive survey on 3D face recognition methods," *Eng. Appl. Artif. Intell.*, vol. 110, Apr. 2022, Art. no. 104669.
- [3] A. Mertan, D. J. Duff, and G. Unal, "Single image depth estimation: An overview," *Digit. Signal Process.*, vol. 123, Apr. 2022, Art. no. 103441.
- [4] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using Laplacian pyramid-based depth residuals," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4381–4393, Nov. 2021.
- [5] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.
- [6] D. Kim, W. Ka, P. Ahn, D. Joo, S. Chun, and J. Kim, "Global-local path networks for monocular depth estimation with vertical CutDepth," 2022, *arXiv:2201.07436*.
- [7] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4009–4018.
- [8] F. Khan, F. Ma, W. Shariff, S. Basak, and P. Corcoran, "Towards monocular neural facial depth estimation: Past, present, and future," *IEEE Access*, vol. 10, pp. 29589–29611, 2022.
- [9] F. Zhang, N. Liu, Y. Hu, and F. Duan, "MFFNet: Single facial depth map refinement using multi-level feature fusion," *Signal Process., Image Commun.*, vol. 103, Apr. 2022, Art. no. 116649.
- [10] J. S. Katroliia, B. Mirbach, A. El-Sherif, H. Feld, J. Rambach, and D. Stricker, "TICaM: A time-of-flight in-car cabin monitoring dataset," 2021, *arXiv:2103.11719*.
- [11] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–16.
- [12] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.
- [13] L. Huynh, P. Nguyen-Ha, J. Matas, E. Rahtu, and J. Heikkil, "Guiding monocular depth estimation using depth-attention volume," in *Proc. 16th Eur. Conf. Comput. Vis.*, vol. 2328. Cham, Switzerland: Springer, 2020, pp. 581–597.
- [14] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," 2019, *arXiv:1907.10326*.
- [15] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.
- [16] R. Dovgand and R. Basri, "Statistical symmetric shape from shading for 3D structure recovery of faces," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2004, pp. 99–113.
- [17] W. A. P. Smith and E. R. Hancock, "Recovering facial shape using a statistical model of surface normal direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1914–1930, Dec. 2006.
- [18] W. Y. Zhao and R. Chellappa, "Symmetric shape-from-shading using self-ratio image," *Int. J. Comput. Vis.*, vol. 45, pp. 55–75, Oct. 2001.
- [19] Q. Jin, J. Zhao, and Y. Zhang, "Facial feature extraction with a depth AAM algorithm," in *Proc. 9th Int. Conf. Fuzzy Syst. Knowl. Discovery*, May 2012, pp. 1792–1796.
- [20] C. Jordan, "Feature extraction from depth maps for object recognition," *Tech. Rep.*, 2013.
- [21] A. T. Arslan and E. Seke, "Face depth estimation with conditional generative adversarial networks," *IEEE Access*, vol. 7, pp. 23222–23231, 2019.
- [22] D. Kong, Y. Yang, Y.-X. Liu, M. Li, and H. Jia, "Effective 3D face depth estimation from a single 2D face image," in *Proc. 16th Int. Symp. Commun. Inf. Technol. (ISCIT)*, Sep. 2016, pp. 221–230.
- [23] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
- [24] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.
- [25] S. Wang, Z. Cheng, X. Deng, L. Chang, F. Duan, and K. Lu, "Leveraging 3D blendshape for facial expression recognition using CNN," *Sci. China Inf. Sci.*, vol. 63, no. 2, Feb. 2020, Art. no. 120114.
- [26] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2D face recognition via discriminative face depth estimation," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 140–147.
- [27] J. R. A. Moniz, C. Beckham, S. Rajotte, S. Honari, and C. Pal, "Unsupervised depth estimation, 3D face rotation and replacement," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–14.
- [28] A. T. Baby, A. Andrews, A. Dinesh, A. Joseph, and V. K. Anjusree, "Face depth estimation and 3D reconstruction," in *Proc. Adv. Comput. Commun. Technol. High Perform. Appl. (ACCTHPA)*, Jul. 2020, pp. 125–132.
- [29] M.-T. Chiu, H.-Y. Cheng, C.-Y. Wang, and S.-H. Lai, "High-accuracy RGB-D face recognition via segmentation-aware face depth estimation and mask-guided attention network," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Dec. 2021, pp. 1–8.
- [30] J. Chen, S. Niu, X. Gao, S. Li, and J. Dong, "SA-UNet for face anti-spoofing with depth estimation," in *Proc. 13th Int. Conf. Graph. Image Process. (ICGIP)*, Feb. 2022, pp. 1–6.
- [31] F. Khan, S. Basak, H. Javidnia, M. Schukat, and P. Corcoran, "High-accuracy facial depth models derived from 3D synthetic data," in *Proc. 31st Irish Signals Syst. Conf. (ISSC)*, Jun. 2020, pp. 1–5.
- [32] F. Khan, S. Basak, and P. Corcoran, "Accurate 2D facial depth models derived from a 3D synthetic dataset," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2021, pp. 1–6, doi: [10.1109/ICCE50685.2021.9427595](https://doi.org/10.1109/ICCE50685.2021.9427595).
- [33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [34] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," 2018, *arXiv:1812.11941*.
- [35] F. Yuan, Z. Zhang, and Z. Fang, "An effective CNN and transformer complementary network for medical image segmentation," *Pattern Recognit.*, vol. 136, Apr. 2023, Art. no. 109228.
- [36] F. Khan, S. Hussain, S. Basak, J. Lemley, and P. Corcoran, "An efficient encoder-decoder model for portrait depth estimation from single images trained on pixel-accurate synthetic data," *Neural Netw.*, vol. 142, pp. 479–491, Oct. 2021.
- [37] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
- [38] G. Borghi, M. Venturrelli, R. Vezzani, and R. Cucchiara, "POSEidon: Face-from-depth for driver pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4661–4670.

- [39] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 11, pp. 1534–1548, Nov. 2014.
- [40] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," in *Proc. Joint Pattern Recognit. Symp.*, vol. 33. Berlin, Germany: Springer, 2011, pp. 101–110.
- [41] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," 2018, *arXiv:1801.09847*.
- [42] A. A. Abdelhamid, E.-S.-M. El-Kenawy, B. Alotaibi, G. M. Amer, M. Y. Abdelkader, A. Ibrahim, and M. M. Eid, "Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm," *IEEE Access*, vol. 10, pp. 49265–49284, 2022, doi: [10.1109/ACCESS.2022.3172954](https://doi.org/10.1109/ACCESS.2022.3172954).
- [43] N. Lopac, F. Hrzic, I. P. Vuksanovic, and J. Lerga, "Detection of non-stationary GW signals in high noise from Cohen's class of time–frequency representations using deep learning," *IEEE Access*, vol. 10, pp. 2408–2428, 2022, doi: [10.1109/ACCESS.2021.3139850](https://doi.org/10.1109/ACCESS.2021.3139850).
- [44] M. A. Farooq, P. Corcoran, C. Rotariu, and W. Shariff, "Object detection in thermal spectrum for advanced driver-assistance systems (ADAS)," *IEEE Access*, vol. 9, pp. 156465–156481, 2021, doi: [10.1109/ACCESS.2021.3129150](https://doi.org/10.1109/ACCESS.2021.3129150).
- [45] M. Valizadeh and S. J. Wolff, "Convolutional neural network applications in additive manufacturing: A review," *Adv. Ind. Manuf. Eng.*, vol. 4, May 2022, Art. no. 100072.



FAISAL KHAN received the B.S. degree in mathematics from the University of Malakand Chankdara, Lower Dir, Pakistan, in 2015, and the M.Phil. degree in mathematics from Hazara University, Mansehra, Pakistan, in 2017. He is currently pursuing the Ph.D. degree with the National University of Ireland Galway (NUIG). He is with FotoNation/Xperi. His research interest includes machine learning using deep neural networks for tasks related to computer vision, including depth estimation and 3-D reconstruction.



WASEEM SHARIFF received the B.E. degree in computer science from the Nagarjuna College of Engineering and Technology (NCET), in 2019, and the M.S. degree in computer science, specializing in artificial intelligence from the National University of Ireland Galway (NUIG), in 2020, where he is currently pursuing the Ph.D. degree (IRC). He is a Research Engineer with Xperi Inc. His research interests include machine learning for computer vision applications, with a particular

emphasis on automotive in-cabin monitoring applications.



MUHAMMAD ALI FAROOQ received the B.E. degree in electronic engineering from Iqra University, in 2012, the M.S. degree in electrical control engineering from the National University of Sciences and Technology (NUST), in 2017, and the Ph.D. degree from the National University of Ireland Galway (NUIG). He is currently a Postdoctoral Researcher with NUIG. Moreover, he is a Machine Learning Research Intern with Xperi Corporation. His Ph.D. research program was funded through the prestigious H2020 European Union (EU) Scholarship. His research interests include machine vision, computer vision, video analytics, machine learning, thermal imaging, and sensor fusion.



SHUBHAJIT BASAK received the B.Tech. degree in electronics and communication engineering from the West Bengal University of Technology, India, in 2011, and the M.Sc. degree in computer science from the National University of Ireland Galway, Ireland, in 2018, where he is currently pursuing the Ph.D. degree in computer science. He has more than six years of industrial experience as a software development professional. He is with FotoNation/Xperi. His research interest includes deep learning tasks related to computer vision.



PETER CORCORAN (Fellow, IEEE) is the Personal Chair of electronic engineering with the College of Science and Engineering, National University of Ireland Galway. He was a Co-Founder of several start-up companies, notably FotoNation (currently the Imaging Division of Xperi Corporation). He has over 600 technical publications and patents, over 100 peer-reviewed journal articles, 120 international conference papers, and a co-inventor of more than 300 granted U.S. patents.

He is currently an IEEE Fellow recognized for his contributions to digital camera technologies, particularly in-camera redevye correction and facial detection. He is a member of the IEEE Consumer Electronics Society for over 25 years. He is the Editor-in-Chief and the Founding Editor of *IEEE Consumer Electronics Magazine*.

...