

Received February 28, 2022, accepted March 9, 2022, date of publication March 11, 2022, date of current version March 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158950

# Towards Monocular Neural Facial Depth Estimation: Past, Present, and Future

FAISAL KHAN<sup>1</sup>, MUHAMMAD ALI FAROOQ<sup>1</sup>, WASEEM SHARIF<sup>1</sup>,  
SHUBHAJIT BASAK<sup>2</sup>, (Graduate Student Member, IEEE),  
AND PETER CORCORAN<sup>1</sup>, (Fellow, IEEE)

<sup>1</sup>Department of Electronic Engineering, College of Science and Engineering, National University of Ireland Galway, Galway, H91 TK33 Ireland

<sup>2</sup>School of Computer Science, National University of Ireland Galway, Galway, H91 TK33 Ireland

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

This work was supported in part by the College of Science and Engineering, National University of Ireland Galway, Galway, Ireland; in part by Xperi Galway, Galway; and in part by the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant 18/CRT/6224.

**ABSTRACT** This article contains all of the information needed to conduct a study on monocular facial depth estimation problems. A brief literature review and applications on facial depth map research were offered first, followed by a comprehensive evaluation of publicly available facial depth datasets and widely used loss functions. The key properties and characteristics of each facial depth map dataset are described and evaluated. Furthermore, facial depth maps loss functions are briefly discussed, which will make it easier to train neural facial depth models on a variety of datasets for both short- and long-range depth maps. The network's design and components are essential, but its effectiveness is largely determined by how it is trained, which necessitates a large dataset and a suitable loss function. Implementation details of how neural depth networks work and their corresponding evaluation matrices are presented and explained. In addition, an SoA neural model for facial depth estimation is proposed, along with a detailed comparison evaluation and, where feasible, direct comparison of facial depth estimation methods to serve as a foundation for a proposed model that is utilized. The model employed shows better performance compared with current state-of-the-art methods when tested across four datasets. The new loss function used in the proposed method helps the network to learn the facial regions resulting in an accurate depth prediction. The network is trained on synthetic human facial depth datasets whereas for validation purposes real as well as synthetic facial images are used. The results prove that the trained network outperforms current state-of-the-art networks performances, thus setting up a new baseline method for facial depth estimations.

**INDEX TERMS** Facial depth datasets, loss functions, neural depth estimation, empirical and systematic evaluation.

## I. INTRODUCTION

The process of obtaining 3D information from a 2D frame is known as depth estimation. Depth estimation is used in diversified computer vision applications such as augmented reality, posture estimation, 3D reconstruction, object detection and recognition, semantic segmentation and -human-machine interaction, weather forecast, and autonomous vehicles. The ground truth depth information used to estimate depth is beneficial for developing reliable navigation systems for intelligent vehicles, environmental reconstruction, and image

interpretation to understand the objects in the image and the scene behind them.

Face depth estimation is a challenging subject that has been explored in conjunction with face motion [1], facial analysis, and facial recognition [2], [3]. Many methods for estimating face depth have been presented in recent years, notably 3D from stereo replicating [4], 3D morphable model-based methods [5], [6], shape from shading (SfS) [5], [6], shape from motion techniques (SfM) [6], [7], and statistical techniques [8], [9]. Due to the facial symmetry of facial areas, the stereo matching procedure for face depth estimation is more complicated (regardless of utilizing the local or global technique), particularly when the system is binocular and therefore only one stereo pair is used. Stereo matching

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li<sup>1</sup>.

methods can estimate a reasonable depth or disparity map for facial depth estimation, but these approaches are more sophisticated, requiring the use of a local or global procedure. Because of the similarity of the face areas, particularly when using a binocular setup with only one pair of stereo images. All stereo approaches are limited by the similarity characteristics of the facial information. Furthermore, the similarity of the pixels values results in more spikes, holes, and particularly uncertain disparities in the depth map.

The computer vision field has conventionally approached the field of depth maps in a variety of methods, such as with stereo or multi-view cameras [10], [11], structure from motion [12], [13], and depth from light diffusion & shading [14], [15]. The described methods face many difficulties, such as missing pixel values and depth consistency, which result in inconsistencies in depth maps. In addition, the camera calibration, camera setup, and post-processing techniques are computationally expensive and time-consuming. The research community has explored the monocular depth estimation task using only a single image which is much more straightforward and suitable for consumer applications. The credit goes to significant advances in machine learning-based networks [16]–[20]. In the first part of the paper, we have given a detailed evaluation of publicly available facial depth datasets and widely used loss functions in facial depth estimation networks, thus to better understanding the problem of facial depth maps. The key characteristics and properties of the facial depth datasets are presented and compared, followed by the loss functions employed. The implementation specifics of how neural depth networks work, as well as the evaluation matrices that correlate to them, are shown and described. A full comparison evaluation and, where possible, direct comparison of facial depth estimation methods are performed in the second phase of the paper to serve as a foundation for a proposed model that is used. When tested across four datasets, the proposed model outperforms current state-of-the-art approaches. The suggested method's unique loss function aids the network in learning the facial areas, resulting in an accurate depth prediction. The network is trained using synthetic human facial depth datasets, and real and synthetic facial images from four facial depth datasets are used for validation.

### A. RESEARCH CONTRIBUTIONS

Following thorough research over the previous few years, image-based facial depth estimation using deep learning algorithms has demonstrated promising results. However, the field is still in its early stages, and more improvements are expected to address issues and challenges such as data selection for training, generalization to unknown environments, fine-scale depth estimation, reconstruction versus recognition, handling multiple objects in the presence of occlusions, and cluttered backgrounds, data imbalance and how to select an appropriate loss function and neural model for facial depth estimation.

This paper aims to provide all of the key information for conducting a study on monocular facial depth estimation challenges. First, a brief review of the literature and applications of facial depth map research was presented, followed by a detailed analysis of publicly available facial depth datasets and commonly used loss functions. To better understand the facial depth map problem, the facial depth dataset's key characteristics and properties are described and evaluated, followed by the loss functions used. For each dataset, the dataset description, metadata, ground truth, and relevant data (year of publishing, ground truth information, image size, type, objects per image, and several images) are listed systematically. In addition, each loss function is presented in such a way that the research community can select the best loss function for their requirements. The implementation details of how neural depth networks work are demonstrated and explained, as are the evaluation matrices that correspond to them. In the second section of the paper, a complete comparison evaluation and, where possible, direct comparison of facial depth estimation methods are conducted to serve as a foundation for a proposed model that is used. The model outperforms current state-of-the-art techniques when tested across four datasets. The unique loss function of the suggested method supports the network in learning the facial areas, resulting in an accurate depth prediction. The network is trained with synthetic human facial depth datasets and validated with real and synthetic facial images from four facial depth datasets.

### B. CHALLENGES AND DEVELOPMENTS

Monocular facial depth estimation based on deep learning (DL) has been intensively explored and advanced over the last few years. However, still, several limitations need to be addressed. This section covers the major issues and discusses potential directions for monocular facial depth estimation maps research. By utilizing a deep learning network, we can extract many features simultaneously, such as semantic information, optical flow features, and depth features. While semantic segmentation will be incorporated into depth estimation, it will remain a separate module that performs autonomous tasks. Additionally, there are typically numerous sub-networks capable of learning depth estimation, visual odometry, and flow estimation. However, such networks are not adequately connected, which results in a large set of network parameters, which eventually requires an increased memory footprint. How to improve the network's integration is a research direction that is worth exploring as the future direction of this research work.

The quality of the training data has a significant impact on the generalization and reliability of the deep learning model. To increase facial depth estimation accuracy, more data with higher quality and a wider variety of scene types is required. However, the facial depth estimation datasets currently available are quite small, and creating a new dataset is time-intensive and expensive. At the moment, several

researchers generate a large number of images for facial depth estimation using a variety of software, but the quality is inconsistent. A future research goal will be to provide a dataset for monocular facial depth estimation that is compatible with deep learning models.

Realistic environments are frequently complex, having a high amount of moving objects, occlusions, changing light conditions, and changing weather. However, the majority of existing facial depth estimation models assume an optimum environment. Although some researchers have attempted to address dynamic objects and occlusion scenarios and have made considerable progress lately, the problem of improving the facial depth estimation of complicated scenes for real-world applications remains a key future research field.

Facial depth estimation is a challenging stage in the development of practical applications such as augmented reality (AR), virtual reality (VR), robotics and autonomous vehicles. However, the resolution of the estimated facial depth is often limited in most existing facial depth estimation algorithms to maximize computational effectiveness.

The fundamental module of SLAM is image depth estimation, which is deeply connected with commercial applications such as autonomous driving. However, researchers frequently design deeper networks with more parameters and constraints to accomplish depth estimation, which needs more computational cost and hence does not fulfil the real-time requirements of modern applications. Thus, a future research area will be to determine how to use a lighter network for real-time estimation while maintaining prediction accuracy.

The rest of the paper is organized as follows: Section 2 discusses related work in the domain of facial depth estimation, especially related studies, or surveys. Section 3 presents the results of a bibliometric investigation, a thorough examination of depth datasets, and further discusses the most used loss functions. Section 4 presents the implementation details of how facial depth neural networks work followed by some comparative analysis of the facial depth estimation methods. Section 5 presents evaluation matrices and section 6 describes and illustrates the most recent SoA depth estimation model, which is discussed and chosen for facial depth estimation. Section 7 shows the experimental results, discusses the training approach, and compares the trained model to SoA methods in a brief comparison study. Section 8 includes a detailed discussion of the experimental results while section 9 provides the conclusion and future research directions.

## II. RELATED WORKS

Datasets are the foundations for evaluating the behaviour and validating the results of artificial intelligence networks, and they play a critical role in scientific research. Another important building block is to use an appropriate loss function to improve the deep network's training performance. An in-depth analysis of various facial depth datasets is performed, and depth regression loss functions for both short and long-range depth datasets are proposed in the next sections.

This section focuses mostly on related facial depth estimation research and applications.

### A. FACIAL DEPTH ESTIMATION APPLICATIONS

Human face images are among the most common images, and they play an important role in many visual interpretations. Since the facial parts separation in a human face is well-known in human anthropometry, it is possible to find the distance of a human focus from a single image frame with good accuracy provided an understanding of the camera's field-of-view. The research community in today's fast-paced technological environment wants more realistic representations, thus 3D representations of 2D images are becoming increasingly important. These methods are categorized into the following primary categories based on their applications.

#### 1) FEATURE EXTRACTION METHODS

The expressions on people's faces reveal information about individuals. Faces identify people, and one may infer how others are feeling from their expressions. Face feature extraction can help in the improvement of face depth maps tasks. In the realm of computer vision, facial feature depth estimation and 3D reconstruction are popular topics. In computer vision-related applications such as detection and recognition, especially under shifting posture lighting, and expression, 3D information gives significant benefits in overcoming difficulties associated with 2D images (PIE) [14]. Methods have been shown in the SoA to be a potential solution to several of problems in facial depth maps [20]–[25].

#### 2) FEATURE FUSION METHODS

Feature fusion offers a full description of image features' rich internal information, and following dimensionality reduction, compact representations of integrated features can be obtained, resulting in decreased computational complexity and better performance of facial depth maps. 3D reconstruction helps in the resolution of difficulties in 2D images as well as the improvement in performance in a variety of tasks. Several approaches have been offered in the last few years [26]–[34] for facial depth estimation tasks.

#### 3) IMAGE PROCESSING FILTRATION METHODS

For the successful application of depth information, quality is critical. Visually undesirable rendered views are frequently produced when a depth map is distorted by large featureless artefacts. A robust depth image post-filtering technique should be considered for further 3D video transmission. Filtering of depth maps has primarily been studied from the viewpoint of increasing resolution [35]–[37]. There are a variety of post-processing techniques for restoring natural images [38]. Filtering algorithms included Gaussian smoothing and the H.264 in-loop deblocking filter [39], as well as a local polynomial approximation (LPA) [40] and bilateral filtering [41], which use edge-preserving structure information from the colour channel to refine rough depth maps [42].

**TABLE 1. Properties of feature, fusion, and image processing filtration methods.**

Method Category	Methods	Strategy	Category	Descriptions of the main block	Uses
Feature Extraction	[14] [20] [22] [23] [25]	Depth From Shading, Defocus Face Depth CNN Recovering Facial Shape Shape-From-Shading From Depth Maps CNN	DL DL ML ML DL	Light-Field Angular Function Adversarial Networks Surface Normal Direction Symmetric Self-Ratio Images Feature Extractor	Depth Maps Depth Maps Reconstructions Reconstructions Object Recognition
Feature Fusion	[26] [27] [28] [29] [30] [31] [32] [33] [34]	Face Depth CNN Face Depth CNN Autoencoder Single Facial Depth Map Face From Depth Face From Depth Pose 3D Blendshape Learning Feature	DL DL DL DL DL DL DL DL ML	Single Reference Face Shape Multi-Level Feature Fusion Stacked Contractive Autoencoder Multi-Level Feature Fusion Feature Fusion Extractor Feature Fusion Extractor Multi-Level Feature Fusion Feature Fusion Extractor Multi-Level Feature Fusion	3D Face Reconstruction 3D Reconstruction Learning 3D Faces Refinement Driver Pose Estimation Image Super-Resolution Pose Estimation Facial Expression Recognition Aggregation
Image Processing Filtration	[36] [37] [38] [39] [40] [41] [42]	Learning Feature Depth Pointwise Shape-Adaptive Pointwise Shape-Adaptive Local Approximation For Gray And Color Images Fused Deep Representation	ML ML ML ML ML DL DL	Joint Bilateral Multistep Joint Bilateral High-Quality Filtration Filters High-Quality Filtration Bilateral Filtering Light Field	Upsampling Depth Upsampling Denoising And Deblocking Deblocking Signal And Image Processing Signal And Image Processing Face Recognition

Table 1 shows the corresponding methods categorized into feature extraction, feature fusion, and image processing filtration with their respective use cases and strategies involved.

*a: FACIAL DEPTH IN 3D FACE RECOGNITION*

Face recognition (FR) has been used for human identification for ages. With the advances of deep neural networks (DNNs), both face identification (one-to-many) and face verification (one-to-one) have achieved state-of-the-art results. Despite these advances, there are still a few limitations due to external conditions like viewing angles, human appearances like facial expressions, occlusions, scene lightings. To overcome these factors researchers, use other modalities like depth and surface normal. The availability of low-cost RGB-D consumer level sensors like Microsoft Kinect and Intel Real Sense which simultaneously capture depth data of the scene and the colour intensity make these multimodal data more accessible. Depth information can be very useful in FR because it helps to retrieve geometric information of the face in the form of dense 3D points. RGB-D FR can be categorized broadly into two classes – handcrafted feature-based method and deep learning-based methods. Table 2 shows the corresponding details of the listed methods for this subsection.

**B. FACIAL DEPTH FROM STEREO AND MULTI-VIEW**

Using two or more cameras, depth can be derived from stereo or multi-view. A process known as stereo matching is used to produce this map. The primary notion is that triangulation and stereo matching can be used to estimate depth in a variety of applications, including object grasping, collision avoidance,

broadcasting, robotic navigation, and multimedia. The most frequently used methods for measuring face depth from stereo methods are designed on fitting the computed depth to a generalized 3D model [49]–[51]. For facial depth estimation, a passive stereo system for 3D human face reconstruction and recognition at a distance method is introduced [52]. Using a Kinect camera and a face detection algorithm, a method was able to reliably locate the human head and estimate head posture. To locate the detailed facial characteristics, a depth AAM algorithm is designed [53]. In a passive stereo vision system, a method for estimating facial depth is introduced. The method relies on the fast creation of facial disparity maps, which does not necessitate the use of expensive instruments or generic face models. It entails including face attributes in the disparity estimate process to improve 3D face reconstruction [54].

The primary drawbacks of these approaches are the long processing times associated with the fitting phase (due to the high computational complexity) and the need for human setup, as seen in [51]. Another drawback of these approaches is that the generated faces resemble the generic model rather than their model. It’s also particularly sensitive to noise because it calculates curves using the second derivative.

**C. FACIAL DEPTH FROM 2D, MONOCULAR IMAGES**

The monocular depth estimation method uses only a single RGB image as input to predict the depth value of each pixel or infer depth information. The following methods use a monocular depth strategy. Monocular depth maps are simple to set up, especially when it comes to camera calibration, and only require a single image to estimate depth. It can also

**TABLE 2.** Properties of facial recognition depth maps methods.

Methods	Feature Type	Features extracted	Strategy	Method Category	Descriptions of the main block	Uses
[43]	Geometric	Histogram Of Oriented Gradient (HOG)	Random Decision Forest (RDF) Classifier	Feature Extraction DL	Entropy Map	Recognition
[44]	Geometric	Local Binary Patterns (LBP)	Iterative Closest Point (ICP) And	Feature Extraction DL	Discriminant Color Space (DCS)	Depth Maps
[45]	Geometric	Signed Distance Function (SDF)	ICP	Feature Extraction ML	3D Face Model	Depth Maps
[46]	Statistical	Feature Space	CNN	Feature Fusion DL	Autoencoder	Depth Maps
[47]	Spatial	Feature Space	Single Facial Depth	Feature Fusion	CNN VGG	Depth Maps & Recognition
[48]	Spatial and Geometric	Feature Space	Face Recognition Accuracy	Feature Extraction	Surface Normal, Point Cloud;	Recognition & Depth Maps

give a variety of monocular visual cues, such as gradients and texture variations, colour, and defocus, that have previously been underutilized in such systems and can be used even in texture fewer areas. Table 3 shows the corresponding details of the listed methods from this section.

#### D. FACIAL DEPTH THROUGH DOMAIN TRANSLATION

The domain translation which is also known as image translation requires learning a parametric mapping function between two separate domains. Per-pixel classification or regression issues are frequently used to solve image-to-image translation challenges [48]–[62]. Borghi *et al.* [30], [51] suggested a method for computing the appearance of a face based on a standard CNN that combines characteristics of autoencoders and fully connected convolutional networks (FCN). Several recent studies have investigated the image-to-image translation problem by developing a mapping between two frames using conditional generative adversarial networks [52], [63]. Authors in [53] and [64], proposed an approach with the pix2pix model, which synthesizes images from semantic labelling and then reconstructs objects from edges and colourizes images. Aissaoui *et al.* [54], [65] provided a framework of linked GANs that can synthesize pairs of similar images in two separate contexts. This research also focuses on the domain translation problem to create visually attractive facial depth maps with sufficient discriminative information for face recognition.

The authors [66] present a novel framework for learning (1) RGB face parsing, (2) depth face parsing, and (3) RGB-to-depth domain translation together for facial depth maps. In [67], the authors suggest a new Deterministic Conditional GAN that is efficient for face-to-face translation from depth to RGB and is trained on labelled RGB-D face datasets. Whereas the network cannot reconstruct the exact somatic attributes of unknown focus on the individual, it can

reconstruct plausible faces which is sufficient for use in various pattern recognition applications. In [68] a method proposes face from depth for head pose estimation on depth images for estimating head and shoulder pose based solely on depth images to create a complete end-to-end system. The proposed method also incorporates head detection and a localization module for facial depth estimation.

#### E. FACIAL DEPTH MAP DENOISING

Two forms of noise which include holes and spikes impact the depth data generated by the face reconstruction process. Pixels with unknown depth values are referred to as holes. During the disparity estimation procedure, the disparity values for these pixels are set to zero. They arise when there is an obstruction or poor light. Spikes are pixels having an incorrect depth estimation. They are mostly caused by incorrect matching and occur inhomogeneous areas where pixels have similar intensity values.

Various approaches for face depth map de-noising have been presented in the literature. These methods are divided into two categories: global and local. To eliminate spikes and fill holes, global approaches apply noise reduction filters to the hole depth image. For this, the median filter is frequently used. Authors in [69] and [70], proposed a Gaussian filter method that works to soften the data and eliminate spikes in the z-coordinate. To eliminate spikes, fill tiny gaps, and smooth the data, the authors in [71] utilized three median filters with different variances. For minor noises, these types of filters can produce optimal results. However, if the noisy region is big, these filters will not be able to remove the noise; instead, they will just modify the pixel values by their surrounding pixels.

In [49] by processing the data row by row, with the first and last non-zero pixels in each row being chosen by a sweep of the depth images. This procedure is continued until no

TABLE 3. Properties of facial depth from 2D monocular images methods.

Methods	Feature Type	Features extracted	Strategy	Method Category	Descriptions of the main block	Uses
[26]	Geometric	Single Reference Face Shape	Constrained Independent Component Analysis	Feature Extraction DL	3D Face Model	3D Face Reconstruction
[9]	Spatial & Geometric	Constrained Independent Component Analysis	The Rotation and Translation Process	Feature Extraction DL	Discriminant Color Space (DCS)	3D Face Reconstruction
[7]	Geometric	Similarity Transform & Feature Space	Deep Learning	Feature Extraction	3D Face Model	Depth Maps & 3D Face Reconstruction
[55]	Statistical	End-To-End Learning	Uses Single-View Depth and Multi-View Pose Networks	Feature Fusion	CNN Models Combined	Depth Maps
[56]	Spatial & Geometric	Canonical Correlation Analysis Surface Depth.	Surface Depth	Feature Extraction	Face Color Texture And Surface Depth	Face Depth Maps
[57]	Spatial & Geometric	Feature Points, Feature Space	Feature Points Similarity Analysis	Feature Extraction DL	Extracted To Form The 2D-3D	3D Face Reconstruction
[58]	Geometric	Recovering The Depth	Uses A Cascaded FCN And CNN Architecture	Feature Extraction	CNN Models Combined	Face Depth Estimation
[59]	Spatial & Geometric	Feature Space	Uses A Combination of Loss Function	Feature Extraction	CNN Encoder-Decoder	Face Depth Estimation

more pixels are produced. The filling process usually involves utilizing an interpolation technique or a local median filter after determining the hole’s boundaries. This method is more accurate than the global method since it just processes noises and leaves the non-noisy data alone. Since holes have a known value (zero or undefined), it can only handle those; spikes, on the other hand, have a random value, therefore it can’t be used to eliminate them.

The authors [72] suggested an edge-guided deep neural network for the super-resolution of a single facial depth map. It is divided into two sub-networks: edge prediction and depth reconstruction. The edge prediction sub-network generates an edge guidance map that is used to guide the depth reconstruction sub-network in recovering sharp edges and fine constructions. Jovanov *et al.* [73] proposes a time-of-flight depth camera-specific wavelet-based depth video denoising approach based on multi hypothesis motion estimation for facial depth maps. In [74] authors proposed a method and system for super-solving and recovering the facial depth maps. The main idea of this approach is to use a learning-based technique to gather reliable face priors from a high-quality facial depth map to improve the depth images.

### III. PUBLICLY AVAILABLE FACIAL DEPTH ESTIMATION DATASETS AND LOSS FUNCTIONS

This section provides an overview of the most commonly used facial image depth datasets, including their respective descriptions in tabular form.

There are several useful datasets available for training depth estimation methods both multi-view and monocular images of human faces. The collection’s general data contains information on the number of objects, scenarios, and RGB and depth images. Among the numerous types of data

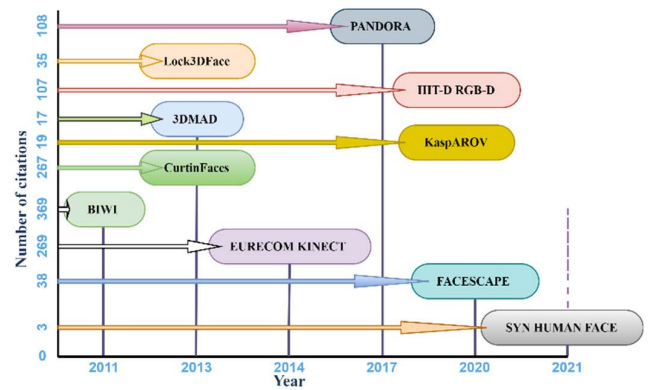


FIGURE 1. The number of facial depth datasets that are publicly available each year.

contained within every dataset, the ground truth contains depth, mesh, cameras trajectories, videos, positions, point cloud, semantics label, trajectories, and dense multi-class labelling. As the field of face image depth estimation research grows in popularity, more work is being put into creating higher and additional informative depth maps datasets. Fig. 1 shows the number of new publicly available facial depth maps datasets and their corresponding number of citations becoming available each year over the period for the last ten (10) years. Table 4-6 tabulates a comparison analysis for the data existing in each dataset.

#### A. FACIAL AND POSE DEPTH DATASETS

The depth camera sensor should be capable of faster human-skeletal tracking in addition to being a low-cost camera sensor

that outputs both RGB and depth information. This kind of tracking can provide the precise position of human body joints throughout a period, making comprehensive human behaviour investigations easier and quicker. As a consequence, there has been a lot of interest in inferring human faces from depth images and synthesizing depth and RGB images. Several new facial depths maps datasets have been generated in recent years to assist in the confirmation of humanoid facemask action analysis methods. The details of these datasets are provided in the following section.

#### 1) BIWI

This dataset [75] comprises 15K images of 20 different subjects which included 6 female subjects and 14 male subjects (4 people were recorded twice). Moreover, this dataset provides the depth image of  $640 \times 480$  pixels resolution, the corresponding visible image of  $640 \times 480$  pixels size, and lastly, it also offers the annotation for every image. The depth data is captured using a Kinect v1 sensor. The dataset consist of the head poses with the range of around  $\pm 75$  degrees yaw and  $\pm 60$  degrees pitch. The overall dataset includes the head's 3D location and rotation as the ground truth data.

#### 2) EURECOM KINECT FACE

This dataset provides multimodal facial data of 52 subjects among which 14 are female, and 38 are male subjects. Eurecom Kinect Face dataset [76] incorporates the depth data which is acquired from Kinect v1 sensor. This data was gathered at different times in the form of two-fold intervals with an average time gap of half month. The recorded data in two different intervals provides the facial frames of each subject in nine situations with various lighting and occlusion conditions and facial expressions which include a neutral face and smiling face.

The provided data incorporates facial data with open mouth, and different occlusions such that strong illumination, eyes occlusion by wearing sunglasses, mouth occlusion by covering it with hand, face side occlusion by placing a paper. The overall dataset provides the RGB colour images, the 3D images, and the depth map which is provided in the forms of the bitmap depth image and the text file containing the actual depth levels acquired from the Kinect sensor. The dataset also incorporates six distinct manual facial landmarks positions which comprise of right and left eye, right and left corner of the mouth, the tip of the nose, and the chin.

#### 3) PANDORA

This dataset [30] provides a total of 250K full-resolution RGB, their corresponding depth data, and their annotations are also included in this dataset. The depth data is acquired from a Kinect v2 sensor. The Pandora dataset is frequently used for various computer vision tasks such that head poses estimation, head centre localization, and shoulder pose estimation.

#### 4) FACESCAPE

The FaceScape dataset [78] includes large-scale 3D facial models, parametric models, and multi-view images all are recorded in high-quality. The dataset also provides the subject's age and gender, as well as the camera settings configuration. The dataset is made publicly available for non-commercial research purposes. This dataset is consisting of 3D faces acquired from 938 subjects. The overall data comprises 18,760 textured 3D faces, with 20 distinct facial expressions. The dataset provides topological information in all the 3D models by processing pore-level facial geometry. For rough shapes and intricate geometry, fine 3D facial models can be expressed as a 3D morphable model, it is represented as displacement maps. A unique methodology is proposed that takes advantage of the large-scale and high-accuracy dataset by utilizing a deep neural network to extract expression-specific dynamic characteristics.

#### 5) 3DMAD

The 3D Mask Attack Database [77] (3DMAD) contains 76500 frames of 17 different subjects captured using the Kinect v1 depth sensor. Each frame is made up of a depth image with an image dimension of  $640 \times 480$  pixels – 1 × 11 bits, a matching RGB image with an image dimension of  $640 \times 480$  pixels – 3 × 8 bits, and precisely labelled eye locations (concerning the RGB image). Data is gathered in three distinct sessions for each subject, with each session consisting of five recordings with each recording including 300 frames. The overall data is recorded from the frontal view with neutral expression in controlled environmental conditions. The complete data is gathered in three different sessions. The first two events are for real-world samples, wherein people are recorded for two weeks. A single operator collects 3D mask attacks in the third session (attacker).

#### 6) SYN HUMAN FACE

The SYN Human FACE [59] includes extensive high-quality 3D face models and their corresponding 2D RGB, pixel-accurate ground truth depth images. The suggested framework works as follows: In Character Creator, a collection of virtual human models is built using the real 100 head models. To generate additional data variations, the texture and morphology of the models are modified. These models are then imported to iClone for incorporating the data with five different facial expressions. The mesh, textures, and animation keyframes for the completed iClone models with individual face emotions are then exported in FBX format.

In the next phase head movement (yaw, roll, and pitch) was applied on all the models in Blender to acquire the head pose. The FBX files are then imported and scaled in the Blender world coordinate system. To replicate the real work environment, lights and cameras are included in the scene, whose properties are then adjusted accordingly. The camera sensor near and far clips have been set at 0.01 meters

TABLE 4. Comparison between data representations.

<ul style="list-style-type: none"> <li>❖ <b>RGB:</b> Images of the visible light spectrum in two dimensions.</li> <li>❖ <b>Depth:</b> The term "depth map" refers to a map of per-pixel data that includes depth-related information. The distance to an object at each pixel is specified by a depth map (e.g., distance from the camera).</li> <li>❖ <b>Video:</b> This type of data displays a series of temporally consecutive visual readings.</li> <li>❖ <b>Point cloud:</b> A 3-dimensional shape is represented by a collection of points, each of which has at least one x, y, and z coordinate.</li> <li>❖ <b>Mesh:</b> It's a polygon-based representation of 3-dimensional objects that captures topological and shape surfaces directly.</li> <li>❖ <b>Scene:</b> It's a form of data that are collected in a specific environment, such as a room or various indoor/outdoor scenarios.</li> <li>❖ <b>Semantic:</b> Labels that relate some data to an ontology class (e.g., human, vehicle, etc.).</li> <li>❖ <b>Object:</b> Object properties such as form, and motion are captured in data. appropriate for tasks such as tracking or object categorization.</li> <li>❖ <b>Camera:</b> This information can be used to track the geometrical properties of the camera.</li> <li>❖ <b>Action:</b> This information is made up of videotapes of people performing specified actions.</li> <li>❖ <b>Trajectory:</b> It is a sort of data that records the course of motion or activity taken by a particular object or entity.</li> <li>❖ <b>Pose:</b> data describing human characteristics, such as head position.</li> <li>❖ <b>Texture map:</b> Texture maps are used to produce repeating textures, patterns, and distinctive visual effects on the surfaces of 3D models. These can be utilized to define precise aspects such as hair, clothing, and skin to any 3D models.</li> <li>❖ <b>UV map:</b> A UV map is a flat representation of a 3D model's surface that is used to wrap textures simply. UV unwrapping is the method of creating a UV map. The term U and V relate to the horizontal and vertical axes of the 2D space.</li> </ul>					
DATA TYPE	DIMENSION	SHAPE INFORMATION	MEMORY PROFICIENCY	COMPUTATION PROFICIENCY	USAGE
RGB	2-D	High	Low	Moderate	Images are detected, represented, and shown in electrical devices like televisions and computers.
Depth	2.5-D	High	Low	Moderate	Simulating the impact of dense semi-transparent material in a scene, such as fog, smoke, or significant amounts of water.
Mesh	3-D	Low	High	Moderate	To form shapes with height, width, and depth, 3D meshes use reference points on the X, Y, and Z axes.
Voxel	3-D	High	Moderate	High	Volumetric imaging in medical and landscape representation in games and simulations.
Point cloud	3-D	Moderate	High	High	from construction and engineering to highway planning and self-driving car development.
Octree	3-D	High	Moderate	Moderate	to recursively subdivide a three-dimensional space into eight octants in order to partition it.
TSDf	3-D	Moderate	High	Moderate	based on a hand-held laser line scanner as a fast, precise, and adaptable geometric fusion method in the 3D reconstruction of industrial products.
Stixel	2.5-D	High	Low	Low	Segmentation, Object tracking.
Texture map	3-D	High	High	High	Generate textures, patterns, or special visual effects.
UV map	3-D	High	Moderate	High	Converting a 3D mesh to a 2D space from a 3D model.







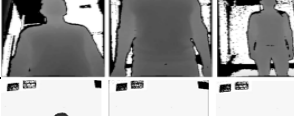
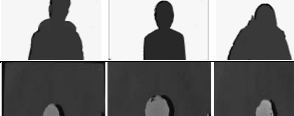
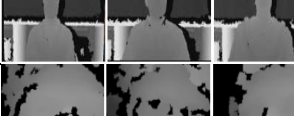
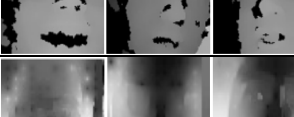

and 5 meters, correspondingly. The sensor size and field of view (FOV) is set to 60 degrees and 36 mm, accordingly. The render layer's RGB and Z-pass outputs are then set up in the compositor to produce the final result. In posture mode, the head and shoulder joints are recognized, the head

mesh has pivoted those bones, and the keyframes are stored to apply the rotation.

Finally, the RGB and depth images are created by rendering all of the keyframes. The matching head position (yaw, pitch, and roll) is produced using the Blender soft-



TABLE 5. Datasets of facial depth, pose, and recognition.

Examples of face images	Dataset	Labelling	Description	camera parameters	APPLICATIONS
	Biwi [75]	3d Position Of The Head And Its Rotation	People Moving Their Heads In Different Directions	Intrinsic + Extrinsic	Automatic Head Pose, Depth, Estimation, Gaze Estimation
	Eure Com Kinectv [76]	Facial Variations, Expressions, Marker Point Positions, Illumination, Occlusion	Performing Various Expressions, Poses	Intrinsic + Extrinsic, Focal Length	Face Recognition, Pose Estimation, Depth Facial Landmark Detection
	3dmd [77]	Spoofing Is Occurring, Eye Positions	3 Different Sessions For All Subjects And Each Session 5 Videos Of 300 Frames Are Captured, Neutral Expression	Intrinsic + Extrinsic	Biometric (Face) Spoofing, Facial Depth Estimation
	Pandora [30]	Head Position And Its Rotation, Features For The Face Verification	People Doing Different Poses In Front Of A Camera Poses	Intrinsic + Extrinsic	Pose, Facial Depth Estimation
	Facescape [78]	Textured 3d Face Models With Pore-Level Geometry, Expressions, Mash, Motion Map, Disparity Map, Texture	Textured 3d Faces, Captured From 938 Subjects And Each With 20 Specific Expressions	Intrinsic + Extrinsic, Focal Length	Predict Elaborate Rig Gable 3d Face Models, Facial Depth Estimation
	Syn Human Face [59]	Expression And Pose, Expressions, Meshes, 3d Position Of The Head And Its Rotation, Lighting	5 Expressions Performed By One Face, Poses, Lighting, Head And Camera Rotation, Translation	Camera Matrix Intrinsic + Extrinsic, Focal Length	Facial Depth Estimation, Pose Estimation
	Baracca Dataset [79]	Measures Of Distance, Age, Weight, Variations, Expressions	In-Car And Outside Views, Human Body Measurements	Intrinsic + Extrinsic	Thermal, Facial Depth Estimation
	Lock3DFace [80]	Changes In Facial Expression, Pose, Occlusion, And Time-Lapse	People Moving Their Heads In Different Directions	Intrinsic + Extrinsic	Pose, Facial Depth Estimation, 3D Face Analysis
	Curtinfaces [81]	Facial Variations, Expressions	Performing Various Expressions, Poses,	Camera Matrix Intrinsic + Extrinsic	Pose, Facial Depth Estimation, Face Recognition
	Iit-D Rgb-D [82]	Head Position And Its Rotation	Performing Various Expressions Poses,	Camera Matrix Intrinsic + Extrinsic	Face Recognition, Facial Depth Estimation
	Kasparov [46]	Variations, Expressions	Poses, Lighting, Head And Camera Rotation	Intrinsic + Extrinsic	Pose, Facial Depth Estimation

ware's python module. For each frame, the RGB images are rendered with a resolution size of  $640 \times 480$  pixels which are then stored in jpg format. Whereas the corresponding depth data is saved in a raw file (.exr format). Moreover, the head poses information for each frame is documented and stored in a text (.txt) file. The rendering process for each 2D frame nearly takes an average time duration of 26.3 seconds which is done using the Cycle Rendering Engine, provided in Blender software which is a type of physically-based path tracer for production rendering. The overall dataset consists

of around 3,500k frames, with around 3.5k 2D frames per person.

The data is stored in a separate folder where each folder contains the data of 100 face models. Each face model's produced RGB images, as well as the resulting depth and head posture, are saved in three separate routes for three different backgrounds: plain, textured, and sophisticated. The synthetic dataset was used to create the sample images, which included ground truth depth images and various backdrops (basic, textured, and sophisticated).

### 7) BARACCA DATASET

The recent interest and growth in depth sensors have supported different methods to instinctively assess the anthropometric measurements, rather than utilising manual procedures and expensive 3D scanners. Normally, the application of depth data is limited due to the lack of depth-based public datasets including accurate anthropometric annotations. As a result, the authors [79] introduced a better dataset, Baracca, that was constructed specifically for the anthropometric measurements and vehicle perspective, including both in-cabin and outside views. This is a type of multimodal dataset that was created with synchronized depth, infrared, thermal, and RGB cameras to meet the needs of the automobile industry. The depth data is recorded using the Pico Zense DCAM710 depth sensor. The spatial resolution of the RGB sensor is  $1920 \times 1080$  pixels, whereas the infrared/depth sensor has a resolution of  $640 \times 480$  pixels. A total of 30 subjects (26 male, and 4 female) took part in the data acquisition process.

### 8) LOCK3DFACE

The Lock3DFace dataset [80] contains 5671 RGBD facial videos from 509 people, each with a unique facial expression, position, occluded, and moments. The database was collected throughout two periods. The very first event's neutral images are used as training examples, while the final three variations are used to create the 3 test procedures for position, occluded, and expressions. All the images from the second run, in all variants, make up a fourth validation set.

### 9) CURTINFACES

CurtinFaces [81] is a well-know RGBD face database that includes over 5000 co-registered RGBD images of 52 participants taken using a Microsoft Kinect. The front left, and right postures are the initial three images for each person. The remaining 49 images include 35 images with 5 different illumination variations and 7 different emotions, as well as 7 distinct positions captured with 7 facial variations. Images with sunglasses and arm occluded are also included in this collection.

### 10) IIIT-D RGB-D

The IIIT-D RGB-D dataset [82] includes 4605 RGBD images from 106 people collected for two periods using a Microsoft Kinect. Each participant was captured with modifications in attitude, emotion, and glasses under typical illumination conditions. The datasets which were before the procedure, which included a 5 cross-validation approach, in the tests set. The head is cropped for each image in the data.

### 11) KASPAROV

The KaspAROV dataset [46], which comprises automatic facial videos from 108 participants is captured by Microsoft Kinect v1 and v2 cameras. Every subject is shown in videos, each shot at a separate time. A total of 432 videos with 117,831 images are included in the dataset. Because the

Kinect v2 sensor data had higher RgbD image registration than the Kinect v1 sensor information.

## B. FACIAL DEPTH ESTIMATION LOSS FUNCTIONS

On the reference depth map, deep learning-based algorithms commonly improve a regression model. The key problem for the SoA approaches in deep regression problems is determining a suitable loss function. Neural networks make use of optimization algorithms.

This error is calculated using the loss function that evaluates how well or badly the model behaves. Neural depth models have been used to estimate depth from one or many 2-D images using a variety of interesting loss functions for depth estimation challenges. This section lists the common loss functions that are used to estimate facial depth maps from one or multi 2D frame images.

### 1) ADVERSARIAL LOSS FUNCTION

The binary categorical cross-entropy loss function, which is used for face depth estimation in adversarial training models [20], [21], is defined as follows:

$$L_{bcc}(\mathbf{y}, r) = -\frac{1}{N} \sum_{i=1}^N [r_i \log y_i + (1 - r_i) \log (1 - y_i)] \quad (1)$$

The discriminator output is subjected to  $y_i = D(I_i)$ , where  $y_i$  is the prediction discriminator for the  $i$ -th input depth map and  $r_i$  is the corresponding ground truth. The goal of the generator model is to create images similar to the GT depth and the discriminator model. The mean squared error (MSE) loss function is used to achieve the first goal.

$$L_{MSE}(y^g, y^d) = \frac{1}{N} \sum_{i=1}^N \|G(y_i^g) - y_i^d\|_2^2 \quad (2)$$

where  $y^g$  and  $y^d$  are the input images and the output depth map. In the second stage of the network, feed created depth images into the discriminator and use the adversarial loss on the discriminator predictions to see if the generated images can trick the discriminator model. Next, while maintaining the discriminator weights constant, back-propagate the gradients up to the generator model input and modify the generator parameters. As a result, the goal of solving the back-propagation problem is to minimize:

$$\hat{\theta}_g = \arg \min_{\theta_g} L_G(y^g, y^d) \quad (3)$$

where  $L_G$  is a balanced sum of two components and can be defined as:

$$L_G(y^g, y^d) = \lambda \cdot L_{MSE}(y^g, y^d) + L_{bcc}(G(y^g), 1) \quad (4)$$

in which  $\lambda$  is a weighting parameter that controls the influence.

### 2) GAN LOSS FUNCTION

The loss function [20], [21] in the GAN-based facial depth model is divided into two parts: 1) Generator Loss: The generator loss is the sigmoid cross-entropy loss of the generated

TABLE 6. Publicly available depth datasets and properties for faces and poses.

DATASET	RGB	DEPTH	VIDEO	POINT-CLOUD	MESH	SCENE	SEMANTIC	OBJECT	CAMERA	ACTION	TRAJECTOR Y	POS E
BIWI [75]	√	√	×	×	×	√	×	×	√	×	×	√
EURECOMKINECT [76]	√	√	×	×	×	√	×	×	√	×	×	√
3DMAD [77]	√	√	×	×	×	√	×	×	√	×	×	×
PANDORA [30]	√	√	×	×	×	√	×	×	√	×	×	×
FACESCAPE [78]	√	√	×	√	√	×	×	×	×	√	√	√
SYN HUMAN FACE [59]	√	√	×	×	√	×	×	√	√	√	×	√
BARACCA DATASET [79]	√	√	×	√	×	×	×	×	√	√	×	√
LOCK3DFACE [80]	√	√	√	×	×	×	×	×	√	√	×	√
CURTINFACES [81]	√	√	×	√	×	×	×	×	√	√	×	√
IIIT-D RGB-D [82]	√	√	×	×	×	×	×	×	√	√	×	√
KASPAROV [46]	√	√	√	×	×	×	×	×	√	√	×	√

No	Dataset Name	Year	Gt	Labeling	Dimension	Objects	Subject/Type	No Images	Diversity	Annotation
1.	BIWI [75]	2011	Depth	Expression, Pose, 2D Skeleton Positions	640 × 480	Multiple	20/realistic	15K	Medium	Real RGB-D
2.	3DMAD [77]	2013	Depth	Expression, Pose, 3D Positions of The Head and its Rotation	640 × 480	Multiple	realistic	76K	Medium	Real RGB-D
3.	CURTINFACES [81]	2013	Depth, Pose	Expression, Pose, 2D Skeleton Positions	640 × 480	Multiple	52/realistic	>5K	High	Real RGB-D
4.	IIIT-D RGB-D [82]	2013	Depth, Pose	Expression, Pose	640 × 480	Multiple	106/realistic	46K	High	Real RGB-D
5.	EURECOM KINECT [76]	2014	Depth	Expression type, Pose, 2D Rotation	256 × 256	Multiple	realistic	20K	Medium	Real RGB-D
6.	LOCK3DFACE [80]	2016	Depth	Expression type, Pose, 3D Position of The Head and Its Rotation	512 × 424	Multiple	509/realistic	>6K	High	Real RGB-D
7.	KASPAROV [46]	2016	Depth	Expression type, Pose, 2D Rotation	64 × 64	Multiple	108/realistic	101K	Medium	Real RGB-D

TABLE 6. (Continued.) Publicly available depth datasets and properties for faces and poses.

8.	PANDORA [30]	2017	Depth	Expression, Pose, 2D Skeleton Positions	256 × 256	Multiple	20/ realistic	11K	High	Real RGB-D
9.	FACESCAPE [78]	2020	2D, 3D Landmarks, Depth	3D Position of The Head and Its Rotation	4096 × 4096	Multiple	938/Extracted	8K	High	Synthetic, 3D, RGB-B
10.	BARACCA DATASET [79]	2020	Depth	Expression, Pose	640×480	Multiple	30/ realistic	>10k	Medium	Real RGB-D
11.	SYN HUMAN FACE [59]	2021	2D, 3D Landmarks, Depth	3D Position of The Head and Its Rotation	640 × 480	Multiple	100/ Extracted	350K	High	Synthetic, 3D, RGB-B

images and an array of ones. The L1 loss function (MAE) is utilized to calculate the absolute difference between the target and generated images. This determines how similar the anticipated image is to the actual image. The following formula can be used to compute the total generator loss:

$$L_{Gen\_loss} = Gan\_loss + \lambda * L1\_loss \tag{5}$$

Here  $\lambda$  is set as 100.

$$MAE = \left(\frac{1}{n}\right) \sum_{i=1}^n |r_i - t_i| \tag{6}$$

where  $r_i$  is the prediction and  $t_i$  are the true value. 2) Discriminator Loss: The discriminator takes real images and generated images as its input. The sigmoid cross-entropy loss of the real images and an array of ones is called real loss. Then the total loss can be calculated by the summation of real loss and the generated loss:

$$T\_loss = Real\_loss + Generated\_loss \tag{7}$$

### 3) STRUCTURAL SIMILARITY (SSIM) LOSS

SSIM [81] is used to determine the perceived differences between the two similar images. ( $L_{SSIM}$ ) represents the loss function for the structural similarity index measure (SSIM) and can be defined as:

$$L_{SSIM}(r, t) = \left(\frac{1 - L_{SSIM}(r, t)}{MaxDepth}\right) \tag{8}$$

### 4) SCALE SHIFT-INVARIANT LOSS

For a single ag image, the scale-shift-invariant loss [81] is defined as

$$L_{SSI}(r, t) = \frac{1}{2N} \sum_i \rho(r, t) \tag{9}$$

where ( $\rho$  is the scale-invariant loss).

### 5) PRE-PIXEL SMOOTHNESS LOSS

Because image gradients commonly have depth inconsistencies, a per-pixel smoothness loss [83] is used in conjunction with the  $L_{SL}$  reprojection loss to make the inverse depth

prediction better. The following formula is used to determine the ( $L_{SL}$ ) loss:

$$L_{SL}(r, t) = \sum_i^N \partial_x dte^{-\partial_x(r,t)} + \partial_y dte^{-\partial_y(r,t)} \tag{10}$$

where N denotes the number of valid pixels,  $\partial d$  denotes the disparity gradient, and  $e^{-\partial x.y(r,t)}$  denotes the edges.

### 6) RECONSTRUCTION LOSS

When training, the network estimates disparity, and the input image is generated using the bilinear samples, utilized to recreate the image. At the local level, the bilinear sampler is completely differentiable and easily integrated into a network. A  $L_{Huber}$  and SSIM is represented as follows: which computes the inconsistencies between both the input image and the regenerated image when coupled as a photometric image reconstruction loss [19].

$$L_R(r, t) = \frac{1}{N} \sum_i \frac{1 - L_{SSIM}(r, t)}{2} + (1 - \alpha)L_{Huber}((r, t)) \tag{11}$$

### 7) SCALE-INVARIANT LOSS

When training the model, depth estimation methods use the GT depth  $y$  and the predicted log depth maps. Scale-invariant loss function [81] ( $L_{SI}$ ) can be represented by ( $L_{SI}$ ) for the depth values and is defined as:

$$L_{SI}(r, t) = \frac{1}{N} \sum_i (\log(r_i) - \log(t_i))^2 - \frac{\lambda}{N} \left(\sum_i \log(r_i) - \log(t_i)\right)^2 \tag{12}$$

where  $\lambda$  refers to the balance factor.

### 8) BERHU LOSS

The OLS estimator is effective in the circumstance of checking for data with outliers or massive errors. Berhu loss, on the other hand, is designed to preserve good attributes in the face of Gaussian noise. Berhu loss function [81] ( $L_{Berhu}$ ) is defined

TABLE 7. Loss functions categorized in terms of the use case applications.

Loss Function	Purpose Of Usage in Terms of Depth Estimation	Other Use Cases
Adversarial Loss Function [20], [21]	The matching feature vectors of distinct identities are linked together to expand the discriminative characteristics between them. The goal is to change the distance between two facial depth image feature vectors and predict the final depth maps.	Segmentation, 3D reconstruction, Synthetic Data generation
Gan Loss Function [22], [23]	This loss function can be used to penalize inter-subject similarities to force the estimated depth image to preserve as much subject discriminative information as feasible.	Segmentation, 3D reconstruction, Synthetic Data generation
Structural Similarity (SSIM) Loss [81]	<ul style="list-style-type: none"> <li>✓ The (Structural Similarity Index) loss function is used with the BerHu loss function to use the input image structure and associated features.</li> <li>✓ The perceptual difference between two similar images is measured by the SSIM loss. Details about structural loss come from relatively adjacent pixels with a deeper connection.</li> <li>✓ These pixels contain vital information about the structure of the visual scene's objects.</li> </ul>	Classification, Regression, Segmentation
Scale Shift-Invariant Loss [39]	<ul style="list-style-type: none"> <li>✓ The loss function with the extra term would create a considerably smaller error because the major issue is to preserve relative depth relationships between pixels.</li> <li>✓ It can also help in a diverse scene such as unknown and inconsistent scales and baselines dataset compatibility. This will allow for data to be trained on from a variety of sensing modalities, including stereo cameras (with potentially unknown calibration), laser scanners, and structured light sensors.</li> </ul>	Regression, Segmentation, Stereo Depth Maps
Pre-Pixel Smoothness Loss	<ul style="list-style-type: none"> <li>✓ This loss function estimates the similarity between the actual and predicted depth map.</li> <li>✓ It also benefits the estimated depth-perceptual map's quality.</li> </ul>	Regression, Segmentation, Stereo Depth Maps
Reconstruction Loss [19]	This loss function can be used to make the projected left-view disparity map equal to the projected right-view disparity map, resulting in more realistic disparity maps.	Segmentation, 3D reconstruction, Synthetic Data generation
Scale-Invariant Loss [39]	<ul style="list-style-type: none"> <li>✓ Regardless of the absolute global size, scale-invariant loss helps in the measurement of relationships between points in the scene.</li> <li>✓ The average deviation between each pixel depth prediction and the ground truth depth is all that is measured.</li> </ul>	Regression, Segmentation, Stereo Depth Maps
Berhu Loss [40]	<ul style="list-style-type: none"> <li>✓ BerHu Loss has an advantage since it uses MSE (or L2) loss to give pixels with greater residuals more weight. At the same time, it allows smaller residuals to have a larger effect on gradients than MAE loss.</li> <li>✓ BerHu's loss function simply combines MAE and MSE, enhancing the whole training process and resulting in more smooth and accurate depth predictions.</li> </ul>	Regression, Segmentation, Stereo Depth Maps
Huber Loss [40]	<ul style="list-style-type: none"> <li>✓ By balancing the MSE and MAE together, the Huber Loss provides the best of both worlds.</li> <li>✓ It is less sensitive to outliers in data and can predict more accurate depth maps.</li> </ul>	Regression, Segmentation, Stereo Depth Maps

as:

$$L_{Berhu}(r, t) = \begin{cases} (r_i - t_i) & \text{if } (r_i - t_i) \leq c, \\ \frac{(r_i - t_i)^2 + c^2}{2c} & \text{if } (r_i - t_i) > c, \end{cases} \quad (13)$$

where  $r_i, t_i$  are ground truth and predicted depth maps.

### 9) HUBER LOSS

MSE is thought to be better at detecting outliers in a dataset, but MAE is expected to be better at preventing them. Data that appear to be outliers, on the other hand, should not be studied, and those points must not be assigned much weight. As a result, the Huber loss function [81] ( $L_{Huber}$ ) is defined as:

$$L_{Huber}(r, t) = \begin{cases} (r_i - t_i) & \text{if } (r_i - t_i) \geq c, \\ \frac{(r_i - t_i)^2 + c^2}{2c} & \text{if } (r_i - t_i) < c, \end{cases} \quad (14)$$

where  $r_i, t_i$  are ground truth and predicted depth maps.

Table 7 shows the loss function categorized according to their use in depth estimation and their respective use case applications.

## IV. IMPLEMENTATION DETAILS OF NEURAL DEPTH ESTIMATION NETWORKS

Convolutional neural networks (CNN) are the form of a learning algorithm for data processing with a uniform grid, such as images, that is intended to acquire provides scalable features from low- to high-level structures efficiently and adaptively. Convolution, pooling, and fully connected layers are the three types of layers (or building blocks) that make up CNNs. Convolution and pooling layers are the initial layers that extract features, while the third, a fully connected layer, transmits these characteristics into the final output, such as classification or multiple regression analysis. A convolution layer is an important part of CNN, which is made up of a stack of mathematical computations like convolution, which is a specific sort of linear operation. Because a feature can appear everywhere in a digital image, image pixels are saved in a two-dimensional (2D) grid, i.e., an array of numbers and a small grid of parameters called the kernel, and an optimizable feature extractor, is implemented at every image position, CNNs are extremely efficient for image analysis. Features extracted can evolve hierarchical structures and progressively

**TABLE 8.** Performance evaluation of monocular depth estimation based deep learning models on IIIT-D RGB-D [82], KASPAROV [46], CURTIN FACES [81], and LOCK3DFACE [80].

REFERENCE	YEAR	NETWORK	DATASETS	PARAMETERS	LAYERS	INPUT/OUTPUT	ACCURACY %
[46]	2016	AUTOENCODER	IIIT-D RGB-D [82]	47M	CNN, FC, SOFTMAX	RGB/DEPTH	98.7
[84]	2014	VGG-16	KASPAROV [46]	32M	CNN, FC, SOFTMAX	RGB/DEPTH	94.4
[85]	2016	RESNET-50	IIIT-D RGB-D [82]	68M	CNN, FC, SOFTMAX	RGB/DEPTH	95.8
[86]	2017	SE-RESNET-50	CURTIN FACES [81]	86M	CNN, FC, SOFTMAX	RGB/DEPTH	97.8
[58]	2018	INCEPTION-V2	LOCK3DFACE [80]	73M	CNN, FC, SOFTMAX	RGB/DEPTH	71.7
[47]	2020	VGG + DEPTH	IIIT-D RGB-D [82]	84M	CNN, FC, SOFTMAX	RGB/DEPTH	99.6

more complicated as one layer passes its results into the next layer. Training is the process of adjusting parameters such as kernels to reduce the disparity between outputs and ground truth labels using optimization algorithms like backpropagation and gradient descent. Fig. 2 illustrates the comprehensive implementation details.

The performance of 2D facial depth estimation has been greatly enhanced because of the use of Deep Learning CNNs. Facial depth maps are learned directly from 2D RGB-D facial images by training deep neural networks on large datasets. Different deep learning models (i.e; VGG, Autoencoder, ResNet, encoder-decoder, inception, DenseNet) are used for facial depth maps which are trained on 2D face depth images. These models typically consist of CNN, FC, SoftMax layers followed by an appropriate loss function that can minimize the errors of the training networks. Weights of the networks are mostly randomly initialized. The datasets can be augmented in several ways (pose augmentation, resolution, transformation, rotation, cropping, and flipping) using a range of images to enlarge training datasets and can achieve better accuracy. Table 8, shows some comparison analysis of the deep learning-based models for facial depth estimation on iiit-d rgb-d [82], kasparov [46], curtin faces [81] and lock3dface [80] datasets. Note that we were unable to compare other qualitative evaluation metrics mentioned in Table 8 due to technical difficulties with publicly available codes and a lack of instructions for these methods

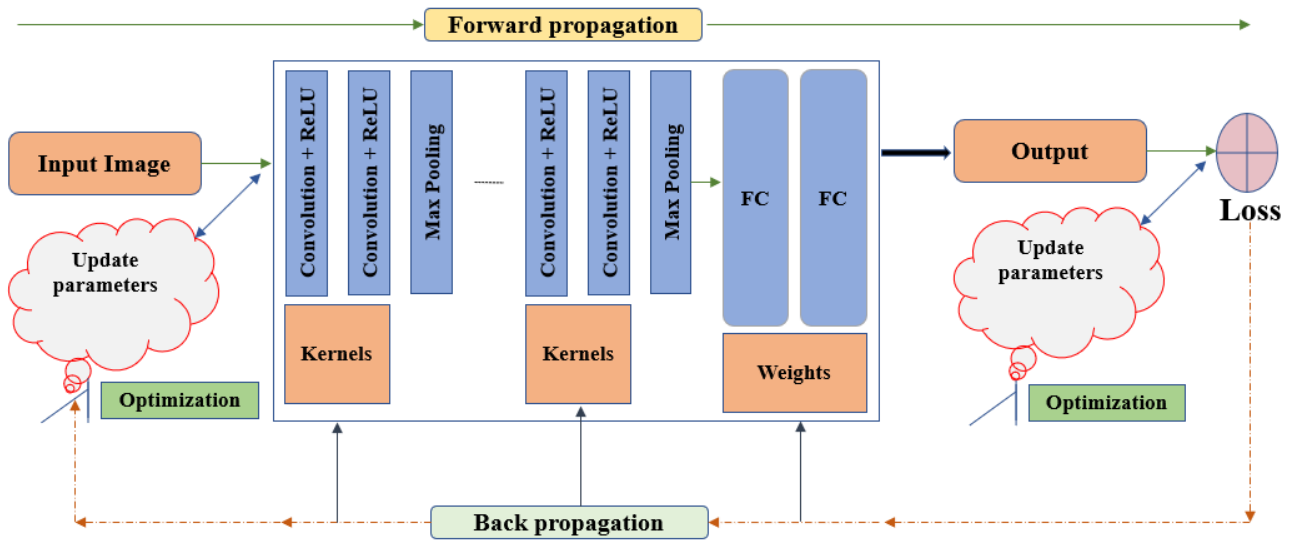
listed in Table 8, and the accuracy results are obtained from their related articles. A CNN-based system has three major components, a training phase, data pre-processing, and model design. To train the model, deep learning-based techniques usually require a significant number of datasets. In CNN-based facial depth maps research, a shortage of large-scale realistic face depth datasets remains an outstanding topic. Because CNN has a lower tolerance for pose changes, suitable data preparation or synthetic data can enhance accuracy before transmitting the data to the model. In addition, selecting an appropriate CNN and loss function are critical.

## V. EVALUATION METRICS FOR FACIAL DEPTH ESTIMATION

The most used quantitative metrics for evaluating the performance of monocular facial depth estimation methods are provided in Table 9. These are not limited to 8 metrics, however, most of the published articles used these quantitative metrics to analyze the performance of the trained depth estimation models.

## VI. FACIAL DEPTH ESTIMATION MODEL

Many consumer applications including robotics, augmented reality and advanced driving monitoring systems can benefit from facial depth estimation neural depth networks from single images. A methodology for creating depth maps from



**FIGURE 2.** A look at the design of a CNN and how it’s trained for facial depth estimation. Convolution layers, pooling layers (e.g., max-pooling), and fully connected (FC) layers are the building components that make up a CNN. The success of a model with certain kernels and weights is evaluated using a loss function and forward propagation on a training dataset, and learning parameters, such as kernels and weights, are adjusted using the gradient descent process. The term “corrected linear unit” refers to a linear unit that has been rectified.

**TABLE 9.** Quantitative metrics used for performance evaluation of monocular facial depth estimation.

S.No	Quantitative Metrics Name	Formula
1	AbsRel	$\frac{1}{N} \sum \frac{ d_i - d_i^* }{d_i}$
2	RMSE	$\sqrt{\frac{1}{N} \sum  d_i - d_i^* ^2}$
3	RMSE (log)	$\sqrt{\frac{1}{N} \sum  \log d_i - \log d_i^* ^2}$
4	SqRel	$\frac{1}{N} \sum \frac{ d_i - d_i^* ^2}{d_i}$
5	Accuracies	% of $d_i, \max(d_i/g_i) = \delta thr$
6	L1	$\sum_{i=1}^n  y_{true} - y_{predicted} $
7	L2	$\sum_{i=1}^n (y_{true} - y_{predicted})^2$
8	NRMSE	$\sqrt{\frac{1}{N} \sum \frac{ d_i - d_i^* }{d_i^*}}$

where  $d_i$  and  $d_i^*$  are the ground truth and predicted depth at pixel  $i$  and  $N$  is the total number of pixels.

single images of human faces is presented in this section, which utilizes the source face depth and corresponding ground truth depth using neural networks.

Existing facial depth map algorithms may produce depth maps with comparable accuracy, but they suffer from difficulties such as missing values and depth similarities, which result in holes in depth images. As an alternative, the model

used in this study automates the collection of optimal parameters, reducing model complexity during the training process for facial depth estimation.

A recent SoA LapDepth [68] model is chosen to accomplish high-quality facial depth estimation from a single 2D frame. By applying the Laplacian pyramid-based decomposition technique to the decoding process, the suggested method intends to successfully restore local details (i.e., depth boundaries) as well as the global layout of the depth map. The depth residual including local details, which suitably describe depth attributes of different scale-spaces, is created using Laplacian residuals of the input colour image guidance encoded features. To improve the efficiency of this decoding process, the authors [87] introduce weight standardization to the pre-activation convolution block, which greatly helps in estimating depth residuals. First, describe the overall architecture of the proposed decoder for monocular facial depth estimation in this section. The entire decoding procedure will then be detailed, including the influence of weight standardization. Finally, the loss functions utilized to train the model architecture are discussed.

### A. ARCHITECTURE DETAILS

The proposed neural depth network for single image facial depth maps mechanism is provided in this section, as well as the suggested loss function for improving the training process over the training data.

#### 1) ENCODER MODEL

The proposed method’s general architecture is demonstrated in Fig. 3 [87]. The suggested decoder for restoring depth residuals is connected to the pre-trained encoder in the

network. ResNext10 [56] is used in the encoder phase, which has been pre-trained for image classification. The input colour image is compressed as latent information using densely layered convolution blocks on the encoder. The spatial size of such features shrinks to a fraction of the original resolution, but they compactly contain the colour-depth relationship in the embedding space, which is learned from various scene geometries. For the convolution block of the encoder, the authors utilize the Dense ASPP approach [88] with four dilation rates of 3, 6, 12, and 18 to extract more dense contextual information.

The suggested decoder is separated into many Laplacian pyramid branches. One branch, which is in charge of the Laplacian pyramid's topmost level, undertakes decoding work to restore the depth map's global layout. The depth residuals are generated by other branches using latent features led by Laplacian residuals of the input colour image at the matching scale. Using point-wise addition, this depth residual is gradually integrated with the middle depth map, which is the result of the higher level of the Laplacian pyramid. The decoding technique is based on a five-level Laplacian pyramid. All convolution layers in the decoder have a filter size of  $3 \times 3$ .

## B. DECODER MODEL

The laplacian residual of the input colour image is derived in the first phase. For all scaling methods in the suggested methodology, downsampling the initial input image, upsampling, and bilinear interpolation are used. Concatenated features are input into layered convolution blocks, and the output is added pixel-by-pixel. The one-channel output, which is made up of stacked convolution blocks, has the same spatial resolution as the input colour image. It's important to note that input guides the decoding process to precisely restore local characteristics of various size areas, which aids in revealing depth boundaries without distortions. Finally, starting at the top of the Laplacian pyramid, the depth map is gradually recreated. The weight standardization in the pre-activation convolution block, which is the core module of the decoder, is made to produce the decoding process for monocular facial depth estimation more effectively. Because the depth map is reconstructed using an iterative accumulation of depth residuals, it is preferable for the projected depth residual to have a balancing of negative and positive values to estimate depth information reliably and accurately. During backpropagation, which is calculated from each layer of the laplacian pyramid, the decoder is capable of improving the flow of gradient by normalizing them. This is preferable for maintaining the colour-to-depth translation's stability based on residual information. The procedure is anticipated to be able to effectively understand the important connection between colour and depth values for facial images by combining this benefit with the Laplacian pyramid-based decomposition technique.

## C. LOSS FUNCTION

The facial depth estimation task's final goal is to find a function that predicts the depth from an input image. ( $L_{silog}$ )

is the most common loss function that is found in the literature more helpful for depth estimation, The network's trainable parameters are tuned based on the loss function, which employs properly scaling the loss function's range can improve converging and training outputs while putting a stronger focus  $\lambda$  on decreasing error variance, leading in a Silog loss function [89]. ( $L_{silog}$ ) is defined:

$$L_{si}(d_i, d_i^*) = \frac{1}{N} \sum_i^N (\log(d_i) - \log(d_i^*))^2 - \frac{\lambda}{N} \left( \sum_i^N \log(d_i) - \log(d_i^*) \right)^2 \quad (15)$$

where  $\lambda$  is the balance factor and  $N$  is the number of pixels.

By rewriting the equation. 15:

$$L_{silog}(d_i, d_i^*) = \frac{1}{N} \sum_i^N (\log(d_i) - \log(d_i^*)) - \frac{1}{N} \sum_N^i (d_i - d_i^*)^2 + (1 - \lambda) \frac{1}{N} \sum_N^i (d_i - d_i^*)^2 \quad (16)$$

In log space, the combined Silog loss is defined as:

$$L_{silog}(d_i, d_i^*) = \alpha \sqrt{L_{silog}(d_i, d_i^*)} \quad (17)$$

## VII. EXPERIMENTAL RESULTS

The experimental results are presented in this section show how well the proposed model performs. The purpose of these experiments is to see how well synthetic facial depth data can be used to estimate facial depth estimation. A set of SoA depth estimation single image neural networks is used to analyze and compare the human facial depth estimation. Furthermore, the model is first trained on a synthetic human facial depth dataset, after which it is evaluated against four different datasets (Pandora, Eurecom Kinect Face, Biwi Kinect Head Pose, and Synthetic human face datasets) explained in section 3. After that, there is a brief comparison analysis (evaluation results of the SoA to the proposed model) is presented. The experiments show that a model trained on a large and diverse set of facial depth images, along with the appropriate training methods, produce SoA results in a variety of scenarios. The zero-shot cross-dataset transfer technique is used to demonstrate this process.

### A. TRAINING METHODOLOGY

The proposed approach is designed in the PyTorch tool. The suggested decoder's parameters (i.e., the network's weights) are all initialized using the approach described in [88]. The proposed decoder has group normalization in each layer, which is known to be batch size independent. The model is trained on a synthetic human facial depth dataset (described in section 3), which was divided into training and validation sets with 0.8 and 0.2 ratios for facial depth estimation. The network is trained using the Adam optimizer for 50 epochs with a batch size of 6, with power and momentum set to 0.9 and 0.999, respectively. For the encoder and decoder, the weight decaying factor is set to 0.0005 and 0. Using a polynomial



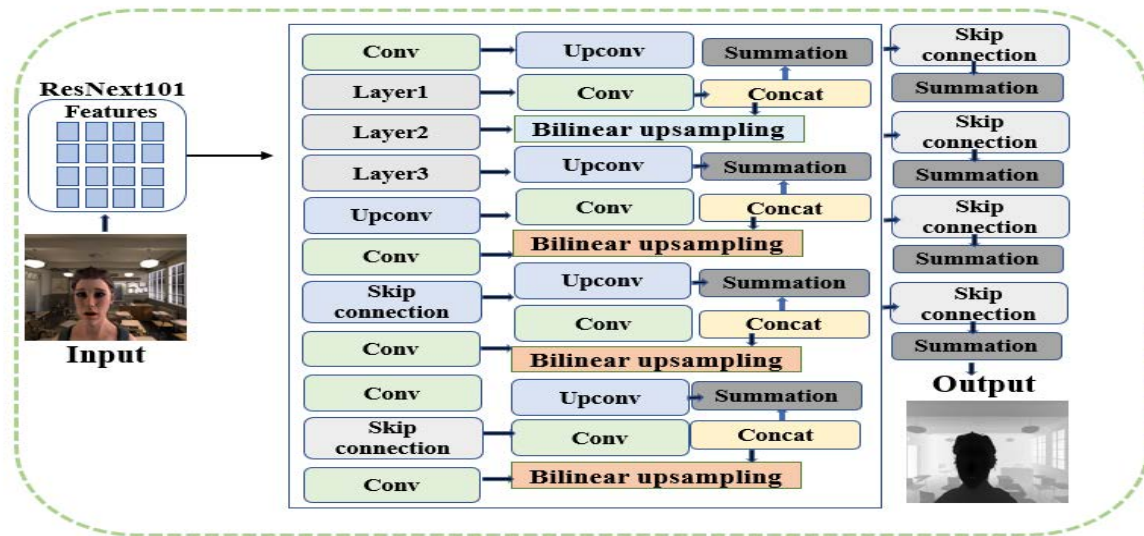


FIGURE 3. The overall architecture of the proposed method for monocular facial depth estimation.

decay with the power of 0.5, the learning rate is first set to  $10^{-4}$  and then gradually decreased until it reaches  $10^{-5}$ . The overall training process is conducted on a machine equipped with two TITAN 1080 GPUs, which takes a time duration of 72 hours. The model has 73M parameters and to avoid overfitting, the online data augmentation method is used in the training process. For the SYN HUMAN FACE dataset, training samples are randomly cropped to  $512 \times 416$  pixels before being randomly rotated in the range of  $[3, 3]$  degrees. With a ratio of 0.5, input images are also horizontally flipped. Furthermore, the scale factor picked from the range of  $[0.9, 1.1]$  is used to alter the brightness, colour, and gamma values of the input colour images.

### B. EXPERIMENTAL DETAILS AND RESULTS

The first phase of this subsection explains the training dataset that was used to train the neural depth model for facial depth estimation. The second part explains the testing and evaluation process used to evaluate the model’s generalization performance. For evaluations, Root Mean Square Error (RMSE), log Root Mean Square Error (RMSE (log)), Absolute Relative difference (AbsRel), Square Relative error (SqRel) and Accuracies are used defined in Table 9. Four test datasets were chosen based on the diversity and accuracy of their ground truth. The model’s performance is compared to existing SoA approaches in the final phase. Table 10 summarizes all of the information from this study’s experiments.

#### 1) MODEL TRAINING DATASET

The synthetic human facial dataset having various variations including camera location, light position, body-pose, facial animations, scene illuminations, and pixel-accurate ground truth depth is used for training the proposed neural depth model for facial depth maps. This dataset is briefly explained

TABLE 10. Information about how experiments have been conducted.

Method	LapDepth [87]
Tools/Software	PyTorch, Open3d
Training Time	72 hours
Input	$512 \times 416$
Output	$512 \times 416$
Type	CNN (Encoder-Decoder)
Optimizer	Adam
Learning Rate	$10^{-5}$
Environment	2×TITAN 1080 GPUs 2.5Ghz Python
Memory	16×2GB
Epochs	50
Parameters	73M

in (section 3-part A subsection 6. Before conducting any experiments, the training data is processed and split into three sets: training set 80%, validation set 20%, and test set 10%, each having its ground truth depth.

#### 2) TEST DATASETS

For comparison purposes, the zero-shot cross-dataset transfer protocol is utilized. The model was trained on a single dataset before being tested on unseen test datasets. The four datasets described in (section 3-part A) were chosen for testing and evaluation (i.e, Pandora, Eurecom Kinect Face, Biwi Kinect Head Pose, and Synthetic human face datasets).

#### 3) MODEL PERFORMANCE EVALUATION

The performance of the facial depth estimation model LapDepth [87] is compared to the SoA models (i.e., MiDaS



**FIGURE 4.** Qualitative results in a sample of the synthetic human facial test dataset that was not used for training or validation. Input RGB images, ground truth images, predicted depth images, predicted depth images (Greys), and predicted depth images are shown from left to right.

[90], DPT [91], and BTS [89]) on the synthetic human facial dataset in Fig. 4 and Table 11. All of the training and testing experiments in this work have been coded and are available on Github. The network achieves SoA results, as shown in Table 11. The proposed model qualitative results against SoA approaches are shown in Fig. 5 and Fig. 6. As shown in Fig. 5, the results demonstrated a details information and consistency, indicating that the proposed chosen approach works better at facial depth estimation. The model outperformed SoA both numerically and qualitatively in tests across a variety of real and synthetic images and set a new SoA for facial depth estimation.

In comparison to other SoA methods, the LapDepth approach performed best in terms of accuracy and depth range, according to the comparison analysis Table 11 and Fig. 6. As shown in Table 11, the network achieved 0.0281 RMSE and 0.9976 threshold accuracy on a synthetic human facial dataset (row 8). For better visualization, the results are shown in the different colour maps. Note that, predicted depth images (Greys) indicate the inverse depth map Fig 4.

As mentioned before the most commonly used quantitative metrics for evaluating the performance of trained monocular facial depth estimation methods are provided in Table 9. Based on the metrics in Table 11 i.e.; RMSE, RMSElog, SqRel, AbsRel, and accuracies one can compare and decide which method performance is better.

The model is compared with the SoA models (i.e.; MiDaS [90], DPT [91], and BTS [89]) for comparison, and the qualitative results are shown in Fig. 5. We were unable to train the techniques (i.e. MiDaS, DPT) from scratch due to unavailability of the training codes and a lack of instructions,

and hence simply fine-tuned the model checkpoint for testing and validation purposes. The method BTS is initially trained on a training dataset before being put to the test on four different datasets. The suggested method has an advantage over the BTS and other SoA methods, as shown in Fig. 5. The model can recover fine details such as facial information and backgrounds since it is trained on pixel-accurate ground truth depth facial data. Pandora, Eurecom Kinect Face, and Biwi Kinect Head Pose are among the datasets that rarely capture those details. It is difficult to learn when training neural depth networks due to a very sparse ground truth depth. It is noticed that the method LapDepth successfully preserves the facial depth information even with complicated geometries as compared to the rest of the SoA approaches. As can be seen in Fig. 6, the results show improved information and consistency, demonstrating that the works were better at depth estimation on real facial depth datasets. The network was not used for training or validation, and the method was exclusively trained on synthetic human facial depth datasets and tested on real datasets. In fig. 5, the results in the 4<sup>th</sup> column predicted depth images (Greys) indicate the inverse depth maps that is originally used by MiDaS [90]. The rest of the comparison results are respectively calculated with the same scale while predicting the depth estimation models.

## VIII. DISCUSSION

The results presented in the previous section are discussed in the following section.

1. The model is trained by using only the Synthetic Human Facial Depth Dataset and evaluated against four different datasets, including the Pandora dataset, Eurecom Kinect Face dataset, Biwi Kinect Head Pose

TABLE 11. Quantitative evaluations on the SNY human face dataset [59].

No.	Methods	AbsRel	SqRel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
1.	DenseDepth-169 [92]	0.0296	0.0096	0.0373	0.0129	0.9890	0.9920	0.9981
2.	ResNet-101 [59]	0.0123	0.0210	0.0306	0.0089	0.9938	0.9965	0.9980
3.	EfficientNet-B0 [93]	0.0145	0.0280	0.0360	0.0154	0.9912	0.9934	0.9978
4.	BTS [89]	0.0165	0.0092	0.0206	0.0102	0.9830	0.9943	0.9956
5.	UNet-simple [94]	0.0103	0.0207	0.0281	0.0089	0.9960	0.9976	0.9987
6.	MiDaS [90]	0.0146	0.0204	0.03560	0.0323	0.9665	0.9902	0.9956
7.	DPT [91]	0.0156	0.0106	0.0394	0.0184	0.9567	0.9646	0.9943
8.	LapDepth [87]	0.0145	0.0041	0.0204	0.3614	0.9545	0.9857	0.99582



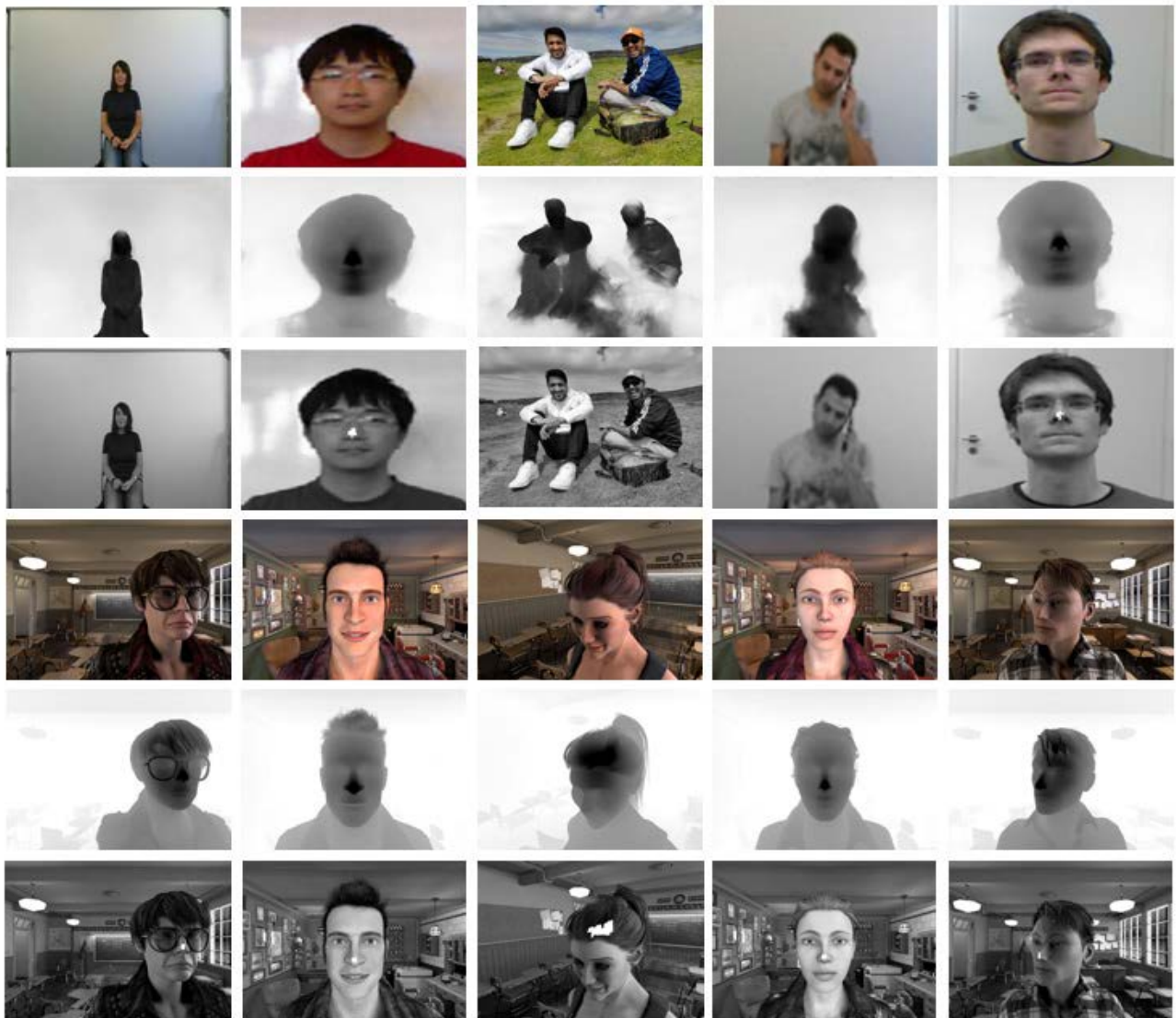
FIGURE 5. From left to right, qualitative results of facial monocular depth estimation algorithms (Input: input RGB images; GT: ground truth images; Ours: LapDepth [87], MiDaS [90], DPT [91], and BTS [89] applied to the Synthetic human facial dataset [59]).

dataset, and the test Synthetic Human Facial Depth Dataset, as well as real images, in the testing phase. The results demonstrate that the trained model outperforms the other SoA approaches MiDaS, DPT, and BTS. It is important to mention that the low size and diversity of the Pandora dataset, Eurecom Kinect Face dataset, Biwi Kinect Head Pose dataset do not perform well on the generalization performance of the studied models, as shown in Fig. 6. Furthermore, most depth GT are error-prone due to practical restrictions in data gathering. The depth GT data is particularly prone to mistakes in these datasets that make it difficult for models to learn robust facial depth information.

2. Synthetic facial data will, of course, lack the same level of detail in terms of skin features as compared to real-world image data. However, considering the numerous advantages of utilizing synthetic data to train

a neural depth model, it acquires comparable accuracy to real-world data as shown in Fig. 6.

- When the new loss function is utilized in the final set of experiments, the model outperforms SoA when the network is trained entirely on synthetic data. As a result, it is rational to assume that employing a scalable loss function and training technique helps in acquiring greater accuracy and facial depth information.
- The model measure how effectively the created faces preserve the individual visual features of the subjects, which requires both high and low-level features to work effectively. The suggested model allows for the maximum test accuracy and outperforms the previous models that have been examined. Based on the results, the model can estimate both high-level and low-level aspects of facial depth maps, resulting in realistic and discriminative results.



**FIGURE 6.** The results of a facial monocular depth estimation method's qualitative evaluation. It demonstrates how to use data from several, independent sources to estimate facial depth in a single view, despite changing and unknown depth range and scale. The method allows for broad generalization across datasets. Input images at the top. Middle: depth maps predicted by the approach provided. Bottom: corresponding point clouds as seen from a different perspective. Open3D [95] was used to render point clouds. Images from the Synthetic human facial dataset, the Pandora dataset, the Eurecom Kinect Face dataset, and the Biwi Kinect Head Pose dataset, as well as a real image of the main authors that were not seen during training.

5. Using the model predicted depth maps, as shown in Fig. 6 (row 3 and 6), the corresponding point clouds can be generated from a different perspective. Many developing visual applications require quick, direct, and exact depth information, which points clouds deliver. To localize and navigate, autonomous technologies such as robots, augmented reality devices, and self-driving cars rely on depth. In high-end smartphones, depth also enables computational photography functions like auto focus and portrait mode, which are especially useful at night when depth is difficult to obtain with traditional cameras but is readily available from a LiDAR.

## IX. CONCLUSION AND FUTURE RESEARCH

This paper investigated the comprehensive details of facial depth datasets and loss functions generated in the field of computer vision for facial depth estimation problems. In various facial depth map tasks based on deep learning networks, publicly available facial depth datasets and facial depth-based loss functions have obtained robust results. The facial depth datasets are utilized in a variety of applications, including person detection and action recognition, face and pose detection, and biomedical applications. Implementation details of how neural depth networks work, as well as their associated evaluation matrices, are presented in this study. In addition to this, SoA neural architecture for facial depth

estimation is proposed, along with a comparison evaluation. The proposed model outperforms current SoA techniques when tested against four different datasets. The proposed method's unique loss function helps the network in learning information aspects more robustly thus providing a detailed prediction. The training is done using synthetic human facial depth datasets, while the evaluation is done with real as well as synthetic facial images. The results prove that the proposed neural model outperforms current SoA networks, thus establishing a new benchmark for facial depth mapping and research aspects. Also, the achieved results presented in this paper can be utilized as a reference for better facial depth estimation model design and validation purposes.

Future research can be focused on developing more robust neural networks, as well as paying more attention to the newly developed facial depth datasets to obtain pixel-accurate ground truth depth maps. Because the currently available datasets have issues, particularly with realistic human faces, they can be employed in a range of real-world applications such as in-cabin driver monitoring, robotics, and 3D face reconstructions if these difficulties are addressed.

Finally, the available SoA depth estimation models can be reconsidered for the prediction of facial depth maps because they are mostly used for indoor and outdoor scene tasks and have not been extensively studied for human faces. They can also be investigated for other tasks such as single view facial recognition and surface normal prediction, 3D reconstructions, and while training on datasets both real and synthetic. The GitHub code is available online and can be found at this URL <https://github.com/khan9048/LapDepth-for-Facial-depth-estimation->.

## REFERENCES

- [1] S.-F. Wang and S.-H. Lai, "Reconstructing 3D face model with associated expression deformation from a single face image via constructing a low-dimensional expression deformation manifold," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2115–2121, Oct. 2011.
- [2] L. Spreeuwers, "Fast and accurate 3D face recognition," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 389–414, Jul. 2011.
- [3] R. Lengagne, P. Fua, and O. Monga, "3D stereo reconstruction of human faces driven by differential constraints," *Image Vis. Comput.*, vol. 18, no. 4, pp. 337–343, Mar. 2000.
- [4] J. Choi, G. Medioni, Y. Lin, L. Silva, O. Regina, M. Pamplona, and T. C. Faltemier, "3D face reconstruction using a single or multiple views," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3959–3962.
- [5] C. K. Chow and S. Y. Yuen, "Recovering shape by shading and stereo under Lambertian shading model," *Int. J. Comput. Vis.*, vol. 85, no. 1, pp. 58–100, Oct. 2009.
- [6] H.-S. Koo and K.-M. Lam, "Recovering the 3D shape and poses of face images based on the similarity transform," *Pattern Recognit. Lett.*, vol. 29, no. 6, pp. 712–723, Apr. 2008.
- [7] Z. L. Sun, K. M. Lam, and Q. W. Gao, "Depth estimation of face images using the nonlinear least-squares model," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 17–30, Jan. 2013.
- [8] J. Fortuna and A. M. Martinez, "Rigid structure from motion from a blind source separation perspective," *Int. J. Comput. Vis.*, vol. 88, no. 3, pp. 404–424, Jul. 2010.
- [9] Z.-L. Sun and K.-M. Lam, "Depth estimation of face images based on the constrained ICA model," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 360–370, Jun. 2011.
- [10] K. Konda and R. Memisevic, "Unsupervised learning of depth and motion," 2013, *arXiv:1312.3429*.
- [11] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun, "Continuous Markov random fields for robust stereo estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 45–58.
- [12] P. Cavestany, A. L. Rodriguez, H. Martinez-Barbera, and T. P. Breckon, "Improved 3D sparse maps for high-performance SFM with low-cost omnidirectional robots," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 4927–4931.
- [13] L. Ding and G. Sharma, "Fusing structure from motion and lidar for dense accurate depth map estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1283–1287.
- [14] M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, "Depth from shading, defocus, and correspondence using light-field angular coherence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1940–1948.
- [15] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Opt. Eng.*, vol. 19, no. 1, 1980, Art. no. 191139.
- [16] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2800–2810.
- [17] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 3, Jan. 2014, pp. 2366–2374.
- [18] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 740–756.
- [19] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6602–6611, doi: [10.1109/CVPR.2017.699](https://doi.org/10.1109/CVPR.2017.699).
- [20] A. T. Arslan and E. Seke, "Face depth estimation with conditional generative adversarial networks," *IEEE Access*, vol. 7, pp. 23222–23231, 2019.
- [21] R. Dovgand and R. Basri, "Statistical symmetric shape from shading for 3D structure recovery of faces," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 99–113.
- [22] W. A. P. Smith and E. R. Hancock, "Recovering facial shape using a statistical model of surface normal direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1914–1930, Dec. 2006.
- [23] W. Y. Zhao and R. Chellappa, "Symmetric shape-from-shading using self-ratio image," *Int. J. Comput. Vis.*, vol. 45, no. 1, pp. 55–75, Oct. 2001.
- [24] Q. Jin, J. Zhao, and Y. Zhang, "Facial feature extraction with a depth AAM algorithm," in *Proc. 9th Int. Conf. Fuzzy Syst. Knowl. Discovery*, May 2012, pp. 1792–1796.
- [25] C. Jordan, "Feature extraction from depth maps for object recognition," Tech. Rep., 2013.
- [26] I. Kemelmacher-Shlizerman and R. Basri, "3D face reconstruction from a single image using a single reference face shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 394–405, Feb. 2011.
- [27] A. T. Baby, A. Andrews, A. Dinesh, A. Joseph, and V. K. Anjusree, "Face depth estimation and 3D reconstruction," in *Proc. Adv. Comput. Commun. Technol. High Perform. Appl. (ACCTHPA)*, Jul. 2020, pp. 125–132, doi: [10.1109/ACCTHPA49271.2020.9213233](https://doi.org/10.1109/ACCTHPA49271.2020.9213233).
- [28] J. Zhang, K. Li, Y. Liang, and N. Li, "Learning 3D faces from 2D images via stacked contractive autoencoder," *Neurocomputing*, vol. 257, pp. 67–78, Sep. 2017.
- [29] F. Zhang, N. Liu, Y. Hu, and F. Duan, "MFFNet: Single facial depth map refinement using multi-level feature fusion," *Signal Process., Image Commun.*, vol. 103, Apr. 2022, Art. no. 116649.
- [30] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "POSEidon: Face-from-depth for driver pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4661–4670.
- [31] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.
- [32] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
- [33] S. Wang, Z. Cheng, X. Deng, L. Chang, F. Duan, and K. Lu, "Leveraging 3D blendshape for facial expression recognition using CNN," *Sci. China Inf. Sci.*, vol. 63, no. 2, Feb. 2020, Art. no. 120114.
- [34] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

- [35] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
- [36] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, Jul. 2007.
- [37] A. K. Riemens, O. P. Gangwal, B. Barenbrug, and R.-P. Berretty, "Multi-step joint bilateral depth upsampling," *Proc. SPIE*, vol. 7257, pp. 192–203, Jan. 2009.
- [38] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images," *IEEE Trans. Image Process.*, vol. 16, no. 5, pp. 1395–1411, May 2007.
- [39] P. List, A. Joch, J. Lainema, G. Bjontegaard, and M. Karczewicz, "Adaptive deblocking filter," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 614–619, Jul. 2003.
- [40] V. Katkovnik, K. Egiazarian, and J. Astola, *Local Approximation Techniques in Signal and Image Processing*, 2006.
- [41] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vis.*, 1998, pp. 839–846.
- [42] C. Angermann, M. Schwab, M. Haltmeier, C. Laubichler, and S. Jónsson, "Unsupervised single-shot depth estimation using perceptual reconstruction," 2022, *arXiv:2201.12170*.
- [43] A. Sepas-Moghaddam, P. L. Correia, K. Nasrollahi, T. B. Moeslund, and F. Pereira, "Light field based face recognition via a fused deep representation," in *Proc. IEEE 28th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2018, pp. 1–6.
- [44] L. Jiang, J. Zhang, and B. Deng, "Robust RGB-D face recognition using attribute-aware loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2552–2566, Oct. 2020.
- [45] G. Mu, D. Huang, G. Hu, J. Sun, and Y. Wang, "Led3D: A lightweight and efficient deep approach to recognizing low-quality 3D faces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5773–5782.
- [46] A. Chowdhury, S. Ghosh, R. Singh, and M. Vatsa, "RGB-D face recognition via learning-based reconstruction," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2016, pp. 1–7.
- [47] H. Zhang, H. Han, J. Cui, S. Shan, and X. Chen, "RGB-D face recognition via deep complementary and common feature learning," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, May 2018, pp. 8–15.
- [48] S. Pini, G. Borghi, R. Vezzani, D. Maltoni, and R. Cucchiara, "A systematic comparison of depth map representations for face recognition," *Sensors*, vol. 21, no. 3, p. 944, Jan. 2021.
- [49] V. Le, H. Tang, L. Cao, and T. S. Huang, "Accurate and efficient reconstruction of 3D faces from stereo images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 4265–4268.
- [50] Y. Zheng, J. Chang, Z. Zheng, and Z. Wang, "3D face reconstruction from stereo: A model based approach," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3, Sep. 2007, pp. III-65.
- [51] J. R. A. Moniz, C. Beckham, S. Rajotte, S. Honari, and C. Pal, "Unsupervised depth estimation, 3D face rotation and replacement," 2018, *arXiv:1803.09202*.
- [52] M. Abdelrahman, A. Ali, S. Elhajian, H. Rara, and A. A. Farag, "A passive stereo system for 3D human face reconstruction and recognition at a distance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 17–22.
- [53] G. Kanojia and S. Raman, "FacialStereo: Facial depth estimation from a stereo pair," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, vol. 3, 2014, pp. 686–691.
- [54] A. Aissaoui, J. Martinet, and C. Djeraba, "Rapid and accurate face depth estimation in passive stereo systems," *Multimedia Tools Appl.*, vol. 72, no. 3, pp. 2413–2438, Oct. 2014.
- [55] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.
- [56] M. Reiter, R. Donner, G. Langs, and H. Bischof, *Estimation of Face Depth Maps From Color Textures Using Canonical Correlation Analysis*, 2006.
- [57] D. Kong, Y. Yang, Y.-X. Liu, M. Li, and H. Jia, "Effective 3D face depth estimation from a single 2D face image," in *Proc. 16th Int. Symp. Commun. Inf. Technol. (ISCIT)*, Sep. 2016, pp. 221–230.
- [58] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2D face recognition via discriminative face depth estimation," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 140–147.
- [59] F. Khan, S. Hussain, S. Basak, J. Lemley, and P. Corcoran, "An efficient encoder–decoder model for portrait depth estimation from single images trained on pixel-accurate synthetic data," *Neural Netw.*, vol. 142, pp. 479–491, Oct. 2021, doi: [10.1016/j.neunet.2021.07.007](https://doi.org/10.1016/j.neunet.2021.07.007).
- [60] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, Jul. 2016.
- [61] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 577–593.
- [62] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," 2016, *arXiv:1609.03126*.
- [63] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
- [64] M. Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 469–477.
- [65] D. Huang, K. Ojui, M. Ardabilian, Y. Wang, and L. Chen, "3D face recognition based on local shape patterns and sparse representation classifier," in *Proc. Int. Conf. Multimedia Modeling*, 2011, pp. 206–216.
- [66] J. Lee, B. Bhattarai, and T.-K. Kim, "Face parsing from RGB and depth using cross-domain mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1501–1510.
- [67] M. Fabbri, G. Borghi, F. Lanzi, R. Vezzani, S. Calderara, and R. Cucchiara, "Domain translation with conditional GANs: From depth to RGB face-to-face," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1355–1360.
- [68] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara, "Face-from-depth for head pose estimation on depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 596–609, Mar. 2020.
- [69] S. Berretti, A. del Bimbo, and P. Pala, "3D face recognition using iso-geodesic stripes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2162–2177, Dec. 2010.
- [70] Y. Wang, J. Liu, and X. Tang, "Robust 3D face recognition by local shape difference boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1858–1870, Oct. 2010.
- [71] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn, "A region ensemble for 3-D face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 62–73, Mar. 2008.
- [72] F. Zhang, N. Liu, L. Chang, F. Duan, and X. Deng, "Edge-guided single facial depth map super-resolution using CNN," *IET Image Process.*, vol. 14, no. 17, pp. 4708–4716, Dec. 2020.
- [73] L. Jovanov, A. Pižurica, and W. Phillips, "Denoising algorithm for the 3D depth map sequences based on multihypothesis motion estimation," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, pp. 1–17, Dec. 2011.
- [74] S. Yang, S. Song, Q. Guo, X. Lu, and J. Liu, "Facial depth map enhancement via neighbor embedding," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1249–1254.
- [75] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," in *Proc. Joint Pattern Recognit. Symp.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 6835, 2011, pp. 101–110, doi: [10.1007/978-3-642-23123-0\\_11](https://doi.org/10.1007/978-3-642-23123-0_11).
- [76] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," *IEEE Trans. Syst., Man, Cybern. A, Syst.*, vol. 44, no. 11, pp. 1534–1548, Nov. 2014.
- [77] N. Erdogmus and S. Marcel, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2013, pp. 1–8.
- [78] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape: A large-scale high quality 3D face dataset and detailed rig-gable 3D face prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 601–610.
- [79] S. Pini, A. D'Eusania, G. Borghi, R. Vezzani, and R. Cucchiara, "Baracca: A multimodal dataset for anthropometric measurements in automotive," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–7.
- [80] J. Zhang, D. Huang, Y. Wang, and J. Sun, "Lock3DFace: A large-scale database of low-cost Kinect 3D faces," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.
- [81] B. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 186–192, doi: [10.1109/WACV.2013.6475017](https://doi.org/10.1109/WACV.2013.6475017).

- [82] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, "On RGB-D face recognition using Kinect," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2013, pp. 1–6.
- [83] M. Carvalho, B. L. Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "On regression losses for deep depth estimation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2915–2919.
- [84] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [85] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [86] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- [87] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using Laplacian pyramid-based depth residuals," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4381–4393, Nov. 2021.
- [88] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.
- [89] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," 2019, *arXiv:1907.10326*.
- [90] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," 2019, *arXiv:1907.01341*.
- [91] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.
- [92] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," 2018, *arXiv:1812.11941*.
- [93] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 10691–10700.
- [94] F. Khan, S. Basak, and P. Corcoran, "Accurate 2D facial depth models derived from a 3D synthetic dataset," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2021, pp. 1–6.
- [95] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," 2018, *arXiv:1801.09847*.



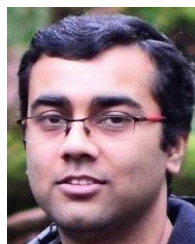
**FAISAL KHAN** received the bachelor's degree in mathematics from the University of Malakand, Chakdara, Pakistan, in 2015, and the master's degree in mathematics from Hazara University Mansehra, Pakistan, in 2017. He is currently pursuing the Ph.D. degree with the National University of Ireland Galway (NUIG). He also works at FotoNation/Xperi. His research focuses on deep neural networks for machine learning applications in computer vision, such as depth estimation and 3-D reconstruction.



**MUHAMMAD ALI FAROOQ** received the B.E. degree in electronic engineering from Iqra University, in 2012, and the M.S. degree in electrical control engineering from the National University of Sciences and Technology (NUST), in 2017. He is currently pursuing the Ph.D. degree with the National University of Ireland Galway (NUIG). His research interests include machine vision, computer vision, video analytics, and sensor fusion. He has won the prestigious H2020 European Union (EU) Scholarship and currently working with NUIG, one of the consortium partners in the Helias (thermal vision augmented awareness) project funded by EU.



**WASEEM SHARIFF** received the B.E. degree in computer science from the Nagarjuna College of Engineering and Technology (NCET), in 2019, and the M.S. degree in computer science, specializing in artificial intelligence, from the National University of Ireland Galway (NUIG), in 2020. He is currently working as a Research Assistant with the NUIG. He is associated with Helias (thermal vision augmented awareness) project. He is also allied with FotoNation/Xperi research team. His research interests include machine learning utilizing deep neural networks for computer vision applications, including working with visible, synthetic data, thermal data, and other bio-sensors.



**SHUBHAJIT BASAK** (Graduate Student Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the West Bengal University of Technology, India, in 2011, and the M.Sc. degree in computer science from the National University of Ireland Galway, Ireland, in 2018. He is currently pursuing the Ph.D. degree in computer science with the National University of Ireland Galway, Ireland. He has more than six years of experience as a software developer in the corporate world. He also works at FotoNation/Xperi. Deep learning tasks relating to computer vision are among his research interests.



**PETER CORCORAN** (Fellow, IEEE) is currently the Personal Chair in electronic engineering with the College of Science and Engineering, National University of Ireland Galway. He was a co-founder of many start-up firms, including FotoNation, which is now part of the Xperi Corporation's Imaging Division. He has over 600 technical publications and patents under his belt, as well as over 100 peer-reviewed journal articles, 120 international conference papers, and is a co-inventor on over 300 granted U.S. patents. For over 25 years, he has been a member of the IEEE Consumer Electronics Society. He has been named an IEEE Fellow for his contributions to digital camera technologies, particularly in-camera red-eye correction and facial recognition. He is the Founding Editor and the Editor-in-Chief of *IEEE Consumer Electronics Magazine*.

...