# Robust Monocular 3D Face Reconstruction Under Challenging Viewing Conditions

Hoda Mohaghegh[a,*], Farid Boussaid[a], Hamid Laga[b], Hossein Rahmani[c] and Mohammed Bennamoun[a]

[a] University of Western Australia
[b] Murdoch University
[c] Lancaster University

ARTICLE INFO

*Keywords*:
Robust 3D Face Reconstruction
Face Texture Refinement
Self-augmentation
Discriminative Face Features
Attribute- aware Loss
Graph Convolutional Networks (GCN).

ABSTRACT

Despite extensive research, 3D face reconstruction from a single image remains an open research problem due to the high degree of variability in pose, occlusions and complex lighting conditions. While deep learning-based methods have achieved great success, they are usually limited to near frontal images and images that are free of occlusions. Also, the lack of diverse training data with 3D annotations considerably limits the performance of such methods. As such, existing methods fail to recover, with high fidelity, the facial details especially when dealing with images captured under extreme conditions. To address this issue, we propose an unsupervised coarse-to-fine framework for the reconstruction of 3D faces with detailed textures. Our core idea is that multiple images of the same person but captured under different viewing conditions should provide the same 3D face. We thus propose to leverage a self-augmentation learning technique to train a model that is robust to diverse variations. In addition, instead of directly employing image pixels, we use a set of discriminative features describing the identity and attributes of the face as input to the refinement module, making the model invariant to viewing conditions. This combination of self-augmentation learning with rich face-related features allows the reconstruction of plausible facial details even under challenging viewing conditions. We train the model end-to-end and in a self-supervised manner, without any 3D annotations, landmarks or identity labels, using a combination of an image-level photometric loss and a perception-level loss that is identity and attribute-aware. We evaluate the proposed approach on CelebA and AFLW2000 datasets, and demonstrate its robustness to appearance variations despite learning from unlabeled images. The qualitative comparisons indicate that our method produces detailed 3D faces even under extreme occlusions, out of plane rotations and noise perturbations where existing state-of-the-art methods often fail. We also quantitatively show that our method outperforms SOTA with more than 30.14%, 9.87% and 11.3% in terms of PSNR, SSIM and IDentity similarity, respectively.

## 1. Introduction

High quality 3D face reconstruction from unconstrained 2D images enables a wide range of computer vision applications, ranging from face recognition [1, 2, 3, 4, 5] and reenactment [6] to virtual and augmented reality [7, 8]. However, inferring 3D geometry and texture from a single image is an ill-posed problem due to the missing 3D information during the projection of a face into the image plane.

Recent years have seen a growing interest in the use of Deep Convolutional Neural Networks (DCNN) for efficient 3D face reconstruction from 2D facial images. These regression-based approaches have demonstrated a significant success due to their powerful representation [9, 10, 11, 12, 13, 14, 15, 16]. Among them, model-based methods which estimate the parameters of a 3D Morphable Model are commonly used to reconstruct 3D facial shapes and texture. However, the fidelity of the texture recovered from the 3DMM coefficients of in-the-wild images is still not sufficiently high. This is because the computed texture from 3DMM cannot capture the fine details of the images that are captured in unconstrained conditions. In particular, these methods are restricted to relatively easy viewing conditions and have difficulties in handling extreme conditions such as occlusions and out-of-plane rotations. Handling occlusions is of high importance since occluders such as sunglasses, hands and long hair can significantly affect the quality of the 3D reconstruction. This can adversely affect the performance of high level tasks such as landmark detection and 3D face recognition, which rely on accurate 3D reconstruction.

In this paper, we propose a novel method for 3D face reconstruction from a single image with the focus on high-fidelity facial texture recovering. The method is required to be robust to occlusions and changes in the viewing conditions. This is achieved through a coarse-to-fine framework in which an initial 3D face is reconstructed, using a 3D Morphable Model (3DMM)-based deep neural network. The reconstructed coarse 3D face is then refined using a novel refinement module, hereinafter referred to as the Fine model, which consists of a 3D mesh-wise refiner and a feature-wise refiner that employ graph convolutions.

In order to reconstruct faces with arbitrary poses and in the presence of occluded regions, the network needs to be trained using a large number of occluded training samples annotated with their corresponding non-occluded ground-truth 3D faces. This is not feasible for in-the-wild images

*Corresponding author
22420326@student.uwa.edu.au ( Hoda Mohaghegh)
ORCID(s): 0000-0002-7086-7975 ( Hoda Mohaghegh)

due to the lack of such paired data. To tackle this problem, we propose a self-augmentation scheme in which, during training, the 3D faces reconstructed by the network are rendered from arbitrary poses, including extreme ones, and with synthetic occlusions. These are then sent back to the partially trained network. In other words, we impose the invariance and robustness of the network to pose variations and (self-)occlusions by pushing the reconstruction of the synthetically synthesized faces to be close to those of the original faces. Also, in contrast to previous methods, we do not propagate the RGB values of the input face image to the vertices of the face mesh in the refinement module. Instead, we incorporate a set of facial attribute-aware and indentity-aware features to make our face representation richer and achieve higher quality reconstruction.

Previous works mostly tackle the 3D face reconstruction task in a supervised manner, which requires a large amount of face images with 3D annotations to reconstruct accurate 3D faces [16, 12, 17, 10, 18, 5]. The major issue with such approaches is the lack of ground truth 3D face data as obtaining scans of face geometry and texture is not easy due to privacy and cost considerations. Even publicly available 3D face datasets, like 300W-LP [17] generally lack diversity in face expression, occlusions and appearance, which jeopardizes the generalization ability of the resulting 3D face regression models. In order to overcome the intrinsic limitations of supervised 3D face reconstruction models, we propose a novel loss function in which we robustify the classic image-level and perception-level losses and combine them to enable robust end-to-end training, in an unsupervised manner without needing any 3D annotations, landmarks or identity labels.

The main contributions of this paper are as follows:

- We develop a fully unsupervised model in which a novel self-augmentation learning offering multiple forms of self-supervision is leveraged to successfully extract plausible facial details even under challenging viewing conditions.
- We propose a rich face representation that integrates face attribute features with commonly-used identity features to capture a high-level understanding of a face. To the best of our knowledge, this is the first attempt towards employing face attribute features for 3D face reconstruction. The proposed set of features can also be used to compute the face attribute distance in our attribute-aware loss, in conjunction with robustified image-level and perception-level losses, giving rise to accurate results.
- Our extensive experiments demonstrate the robustness of the proposed method to challenging viewing conditions including (self)occlusions and noise where existing state-of-the-art (even those trained in a fully supervised fashion) often break down.

The rest of the paper is organized as follows. Section 2 reviews the state-of-the-art in monocular 3D face reconstruction. Our proposed method is presented in Section 3. Section 4 analyzes the performance of the proposed method through extensive experiments. Finally Section 5 concludes the paper.

## 2. Related Work

Various approaches have been proposed to tackle the inherently ill-posed problem of 3D face reconstruction from a single image; see Zollhofer et al. [19] for a detailed survey. Here, we focus on those that are directly relevant to our approach and classify them based on their degree of 3D supervision.

### 2.1. Learning with 3D supervision

There are several learning-based approaches, which train CNNs with 3D annotated training samples to regress 3D faces from images. Model-based methods estimate the parameters of a 3D Morphable Model [20]. For instance, [17, 10, 18] use cascaded CNN structures to regress the 3DMM coefficients, but their approaches are time consuming due to their iterative and multi-stage nature. Tran et al. [13] learned to directly regress 3DMM parameters of a face model by leveraging multiple images of the same subject. The weighted average of the fitted meshes are then used as ground truth to train the network. Liang et al. [5] proposed a mugshot-based 3D face shape reconstruction method implemented by both linear and nonlinear regression and an efficient texture recovery module. Their proposed method generates 3D face images which eliminate texture inconsistency caused by varying illuminations in mugshot images. Tran et al. [21] proposed an encoder-decoder framework to learn a nonlinear 3DMM of facial shape and texture. In comparison to linear models, non-linear models have a higher representation power but still fail to reconstruct facial details due to being trained on the 300W-LP dataset [17], which was generated using a linear 3DMM.

Apart from the aforementioned methods, end-to-end approaches have also been proposed to bypass the 3DMM coefficients regression by directly obtaining 3D faces using CNNs. Jackson et al. [12] use volumetric convolutions to map an input image to full 3D facial structure. While their method is not restricted to a 3DMM space, it introduces heavy computations due to the use of volumetric convolutions. Feng et al. [16] solved simultaneously the problems of 3D face alignment and 3D face reconstruction. They designed a 2D representation called UV position map, which is a 2D image recording the 3D coordinates of a complete facial point cloud in the UV space. To regress this UV position map from a single 2D facial image, an encoder-decoder network was trained with a weighted loss focused on the most discriminative face regions. Despite employing a light weight framework, this method still results in accurately reconstructed 3D faces. However, it has difficulties reconstructing faces taken from extreme viewing angles or faces with occlusions.

All the aforementioned methods require facial images annotated with their corresponding 3D face. Since acquiring 3D ground truth data is laborious and time consuming, a number of works have resorted to synthetic data [9]. These methods, however, may not generalize well to real images because of the large synthetic-realistic domain gap. Sela et al. [22] overcame this problem using an image-to-image translation network followed by geometric deformation and refinement steps to generate detailed 3D faces. Although their method shows good performance on diverse faces under extreme expressions and different viewing angles, its network requires heavy calculation on optimization. Also, the model fails to generalize to cases that significantly deviate from the training domain, e.g., in-the-wild-images with extreme occlusions.

In general, the reconstructions obtained using either volumetric or model-based methods lack details especially when the input images are captured in unconstrained conditions. A number of methods overcome this issue using Generative Adversarial Networks (GANs) to reconstruct [23, 24] or complete [25] the facial texture in the UV space. While they can model, with high fidelity, facial textures, their models require large-scale training sets of high-resolution UV maps, which are difficult to obtain.

Apart from the methods that rely on single still images, another class of methods exploits multiple inputs during training or testing. MGCNet [26] is a model-based method that leverages multi-view geometry consistency to provide reliable constraints on face pose and depth estimation. More recently, Bai et al. [27] presented a method for riggable 3D face reconstruction using an end-to-end trainable network embedded with an in network optimization. However, such a model is still not able to generate results from a single input since it relies on multi-view aggregation during inference.

## 2.2. Learning with weak 3D supervision

To reduce the degree of 3D supervision, Sengupta et al. [28] proposed SfSNet, which learns from a mixture of labeled synthetic and unlabeled real world images. This allows the model to capture high frequency details from real images and low frequency variations from synthetic images using a photometric reconstruction loss. The performance of the model on real world images is limited by the realism and diversity present in the synthetic samples. Zeng et al. [29] proposed a Dense-Fine-Finer framework using a similar training set i.e, a combination of 3D synthetic data, 2D image reconstructed data and fine facial images. The model can reconstruct subtle facial details such as small crow's feet and wrinkles. However, because the original color image is integrated into the input layer of F-Net, the model fails to recover face texture in occluded facial regions. Genova et al. [30] proposed an end-to-end learning scheme using a differentiable rendering layer. By introducing three novel objectives, they trained a regression network using synthetic data and its corresponding 3D parameters in a weakly supervised fashion.

Tu et al. [31] tackle the shortage of 3D annotated training data by using sparse 2D facial landmarks as additional information for the input of the CNN regressor. In addition, the authors introduce self-critic and self-supervision to weakly supervise the training samples using in-the-wild 2D face images only, i.e., without 3D annotations.

## 2.3. Learning without 3D supervision

There is limited work on unsupervised monocular 3D face reconstruction. Deng et al. [32] propose a method to simultaneously predict the 3DMM shape and texture coefficients by exploiting hybrid-level image information. However, their generated textures lack high frequency details since it is produced by a 3DMM model and is inherently limited by the 3DMM texture model. Tewari et al. [15] proposed a Model-based deep convolutional Face Autoencoder (MoFA) to extract semantically meaningful parameters from a single monocular input image in an unsupervised manner. The applied loss is formed by the error between the rendered image and the input. Despite its complexity, the model does not show great generalization capability and has issues with extreme expressions and occlusions. Bas et al. [33] alleviated complexity by using a purely geometric algorithm in which only the shape component of a 3DMM is used to geometrically normalize an image. Tewari et al. [34] further improved MoFA [15] and proposed a multi-level face model that fuses a convolutional encoder with a differentiable renderer and a self-supervised training loss. Their approach achieves improved robustness and quality in terms of geometry and reflectance. However, it struggles with extreme facial expressions.

RingNet [35] introduces a multiple encoder-decoder based architecture optimized for shape consistency across an arbitrary number of input images. It enforces shape consistency between images of the same subjects while enforcing shape inconsistency between images of different subjects. Zhang et al. [36] proposed a Stacked Contractive Autoencoder (SCAE) which learns nonlinear subspaces from both input face images and corresponding 3D faces. Their proposed model learns the mapping from image subspace to 3D face subspace via a one-layer fully connected neural network which is just the realization form of an Autoencoder. In FML [37], Tewari et al. proposed a video-based self-supervised approach for joint multi-frame learning of a face model and a 3D face reconstruction network. The trained network is usable for both monocular and multi-frame face reconstruction. However, if the number of multi-frame images is large, the networks trained on the averaged representations across multiple frames may not work well for a single monocular image, due to the large gap between averaged and single images. Lin et al. [14] used Graph Convolutional Networks (GCN) to reconstruct the detailed colors of the mesh vertices without using any explicit 3D annotations. However, since it directly exploits facial details from the color values of the input image, their model fails to reconstruct faces under occlusions and large poses.

To address the occlusion issues in 3D Face Reconstruction, Egger et al. [38] proposed to combine segmentation and 3D morphable model adaptation. Their occlusion-aware method segments the target image into face and non-face regions, and iteratively adapts the face model and the segmentation to the target image. However, in addition to being slow, this method fails to tackle small-scale or skin-colored occlusions. Egger et al. further extended their previous work [38] by integrating an explicit beard model and a prior during segmentation [39]. Although this explicit modeling of beards improved the quality of fits compared to their previous work [38], but the model still fails to faithfully reconstruct the facial details. Instead of adopting independent per-image optimization for the fitting-segmentation process, Li et al. [40] jointly solved the reconstruction and the segmentation as a learning problem on a larger set of training data. Despite being significantly faster, the segmentation performance relies largely on the generative ability of the face model which can result in unfaithful reconstruction.

Tiwari et al. [41] tackled the issues of occlusions and noise in 3D face reconstruction using a self-supervised robustifying guidance framework, which learns statistical facial coefficients for clean, occluded, and noisy face images simultaneously. However, their method requires the preprocessing of the images which can fail due to several reasons (e.g., the inability of the face detection model to detect the face) and dampen the model's performance. Feng et al. [42] presented Detailed Expression Capture and Animation (DECA), in which the low-dimensional detail latent space makes the fine-scale reconstruction robust to common noise and occlusions. However, their method still does not address extreme occlusions and fails in extreme head pose and lighting.

In this article, we tackle diverse variations (including occlusions, large poses, challenging lighting conditions and appearance variations, e.g., beard and heavy make-up) using the combination of a self-augmentation learning technique and the rich face-related features. We propose to build upon the success of GCNs, but instead of directly using the RGB values of the input image for texture refinement, we take the advantage of a rich discriminative face representation. This representation effectively describes the identity and the traits of facial images such as gender, emotion, age and race.

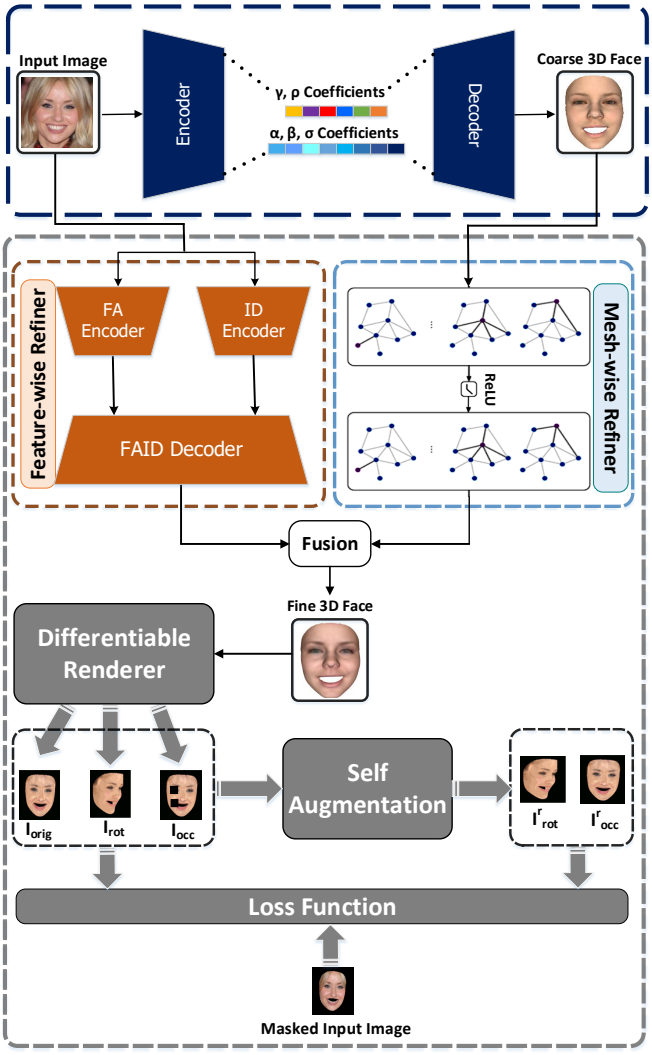


**Figure 1:** Overview of the proposed framework. In Coarse Model (navy blue dashed block), 3DMM coefficients are regressed from the input image and then used to reconstruct the coarse 3D face. The reconstructed face along with the extracted facial features are employed to generate the fine 3D face in Fine Model (gray dashed block).

## 3. Proposed Method

Our goal is to reconstruct, in an unsupervised manner, high fidelity 3D faces with plausible facial details in regions occluded by obstructions. The targeted method should be invariant to viewing conditions. Our key idea is that the 3D face of a person should remain unchanged, in spite of changes in lighting conditions, occlusions or other extrinsic factors. To this end, we introduce (1) a self-augmentation learning scheme embodied in a coarse-to-fine reconstruction framework, and (2) a self-supervised loss functions using a differentiable rendering module to enable robust end-to-end training on a large corpus of in-the-wild face images without the need for 3D annotated ground truth, identity labels or landmarks. In order to represent high fidelity textures and naturally recover unobserved regions, we also propose a refinement step that utilizes a rich face representation describing face attributes and identity instead of propagating RGB values of the input image to the face mesh (See Fig. 1).

In what follows, we first introduce the Coarse Model that regresses the 3DMM coefficients for reconstructing an initial coarse 3D face (Section 3.1) and then describe, in Section 3.2, the refinement process.

### 3.1. Coarse Model

The goal of this module is to reconstruct an initial coarse 3D face from the input face image. To this end, we follow the approach of [32], which regresses from an input image both

the shape and texture coefficients of the 3DMM of a face as well as the rendering parameters (i.e., the illumination and face pose parameters). The estimated 3DMM coefficients are then used to generate the coarse 3D shape and texture of the input face image. In a nutshell, the face shape $S \in \mathbb{R}^{3N}$ that stores the 3D coordinates of N mesh vertices and the texture $T \in \mathbb{R}^{3N}$ that stores the color of N mesh vertices can be represented by a linear combination over a set of PCA basis:

$$S = \overline{S} + B_{id}\alpha + B_{exp}\beta, \tag{1}$$

$$T = \overline{T} + B_t\delta. \tag{2}$$

Here, $B_{id} \in \mathbb{R}^{3N\times80}$, $B_{exp} \in \mathbb{R}^{3N\times64}$, and $B_t \in \mathbb{R}^{3N\times80}$ are, respectively, the PCA basis of identity and expression-dependent variations of the 3D shape of faces. $B_t$ is the PCA basis of texture variations. $\alpha$, $\beta$, and $\delta$ are the corresponding coefficients. $\overline{S} \in \mathbb{R}^{3N}$ and $\overline{T} \in \mathbb{R}^{3N}$ are, respectively, the average face shape and texture. In this work, Basel Face Model (BFM) is adopted for $\overline{S}, B_{id}, \overline{T}$ and $B_t$ and the expression bases $B_{exp}$ are built from the FaceWarehouse [43].

In summary, in this module the vector $(\alpha, \beta, \delta, \gamma, \rho) \in \mathbb{R}^{257}$ is predicted using a ResNet-50 network where $\alpha \in \mathbb{R}^{80}, \beta \in \mathbb{R}^{64}, \delta \in \mathbb{R}^{80}, \gamma \in \mathbb{R}^{27}$ and $\rho \in \mathbb{R}^{6}$ represent the 3DMM shape, expression, texture, lighting and face pose coefficients respectively. Shape, expression and texture coefficients are used in this module to compute the 3D face shape $S$ and texture $T$ using Eq. (1) and Eq. (2). The predicted lighting and face pose parameters are further employed in the Differentiable Renderer module (Section 3.2.3) to render the face mesh to a 2D image.

## 3.2. Fine Model

In order to reconstruct fine facial details, we employ a refinement module, hereinafter referred to as Fine Model. It is composed of a 3D mesh-wise refiner and a feature-wise refiner. More specifically, the initial face reconstructed by the Coarse Model along with the input image are fed into the 3D mesh-wise refiner and feature-wise refiner, respectively. The refiners, trained in a fully unsupervised manner under the self-augmentation technique (see Section 3.2.3), output detailed vertex texture maps (one per refiner), which are then fused to generate the fine-grained 3D face.

### *3.2.1. 3D Mesh-wise Refiner*

We employ a Graph Convolutional Network to generate the detailed face texture from the initial face texture reconstructed in the Coarse Model. A 3D facial texture with $N$ vertices can be represented by an undirected graph $G = (V, A)$ where $V \in \mathbb{R}^{N\times3}$ is a set of vertex colors and $A \in \{0,1\}^{N\times N}$ is the adjacency matrix encoding the connectivity vertices. The normalized graph Laplacian is defined as $L = I_N - D^{-1/2}AD^{-1/2}$, where $I_N$ is the identity matrix and $D \in \mathbb{R}^{N\times N}$ is the diagonal matrix with $D_{ii} = \sum_j A_{ij}$. In Chebyshev Spectral Graph CNN, the spectral convolution for $x_i \in \mathbb{R}^{N\times F_{in}}$, the input with $F_{in}$ features can be defined as [44]:

$$y_i = \sum_{i=1}^{F_{in}} g_{\theta_{i,j}}(L)x_i, \tag{3}$$

where $y_i \in \mathbb{R}^{N\times F_{out}}$ is the output with $F_{out}$ features and $g_\theta(L)$ is the filter parametrized as a truncated Chebyshev polynomial expansion of order K. It is defined as:

$$g_\theta(L) = \sum_{k=0}^{K-1} \theta_k T_k(\widetilde{L}), \tag{4}$$

where $\theta \in \mathbb{R}^K$ is the Chebyshev coefficients vector and $T_k(\widetilde{L}) \in \mathbb{R}^{N\times N}$ is the Chebyshev polynomial of order $k$ computed at a scaled Laplacian $\widetilde{L} = \frac{2L}{\gamma_m ax} - I_N$. $T_k$ can be recursively computed by $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ with the initial $T_0 = 1$ and $T_1 = x$. Based on this concept, in this module, a Graph Convolutional Network, which consists of four spectral residual blocks, takes as input the RGB values of the mesh vertices from a reconstructed coarse face and generates the refined texture for mesh vertices.

### *3.2.2. Feature-wise Refiner*

To further enhance the refined 3D face generated by the 3D mesh-wise refiner module, we propose to extract a set of discriminative features that describe the Face Attribute and IDentity (FAID) of a face, and feed them to our FAID decoder to produce detailed texture for the mesh vertices.

To a large extent, the perception of 3D human faces is correlated with the attributes of the facial appearance including gender, emotion, age, and race. These attributes provide important visual cues for 3D face reconstruction. In other words, faces with the same facial attributes share some common patterns in texture and geometry. For example, Caucasian faces usually have light skin tones, male faces are more likely to grow a beard, and more wrinkles can be found in aged (old) faces. Motivated by this observation, this module extracts a set of Face Attribute (FA) features from the off-the-shelf face attribute classification networks [45] and feed them into the FAID decoder to produce the RGB values for each vertex. We take advantage of a state-of-the-art face recognition network, FaceNet [1], to extract the IDentity (ID) code. Identity features, which are robust to many variations including illumination and pose, are shown to be effective for 3D face reconstruction [30], synthesizing new identities [46], and normalized faces [47]. The discriminative FAID features are also utilized to verify that the reconstructed face resembles the input face in terms of identity and attributes; see Section 3.2.4.

We concatenate the attribute features and the identity features to get a rich face representation as the input to the proposed FAID decoder network. The FAID decoder in the feature-wise refiner is also a graph convolutional network and has the same architecture as the network in the 3D Mesh-wise Refiner module. It takes the Face Attribute and Identity features and outputs the detailed texture for the vertices

of the face mesh. The final vertex texture with details is obtained by simply averaging the outputs of the 3D mesh-refiner and the feature-wise refiner modules.

#### 3.2.3. Self-augmentation learning

We develop a self-augmentation learning scheme to make our model robust against two important challenges in 3D face reconstruction: occlusions and large pose variations. When a model learns to reconstruct 3D faces from a single view, the reconstructed face does not seem accurate when viewed from a different viewpoint. To tackle this issue, we render, using a differentiable renderer, a secondary image called $I_{rot}$ with a random pose and send it back to the partially trained model to generate $I^r_{rot}$. Then, given the input image $I$, the training loss is imposed between $I_{rot}$ and $I^r_{rot}$ and also between $I$ and $I^r_{rot}$ to improve the generalization ability of the model to viewpoint variations.

The intuition behind this scheme is that training on images rendered under various poses helps the model resolve self-occlusions in the training samples. Thus, it improves its robustness compared to single view training. In addition, passing back the reconstructed face through the partially-trained model ensures that the network is able to correctly interpret its own output. The same strategy can also be applied this time by feeding back synthetically occluded rendered images, named $I_{occ}$, to make the model robust to large occlusions caused by accessories, hair, and hand-to-face gestures. In fact, we feed $I_{occ}$ to the partially trained model and force it to generate the same face $I^r_{occ}$ for the same identity. By imposing training loss between $I_{occ}$ and $I^r_{occ}$ and also between $I$ and $I^r_{occ}$, we ensure the image-level and perception-level consistency between the input and the reconstructed image in the presence of occlusions.

We call this strategy *self-augmentation learning* as we encourage the predicted representations of self-augmented data (synthetically occluded faces and multi-view faces) to be as close as possible to those of the original data in an end-to-end fashion. The augmentation function can be designed specifically for any variations that challenge the 3D face reconstruction problem to impose different types of invariances on the representations predicted by our model.

**Differentiable Renderer.** We enable self-supervised training of our model without requiring the training pairs by employing a differentiable rendering module in which a 2D image is rendered from the refined 3D face estimated by the network. To project the estimated 3D face to a 2D face image, we adopt a differentiable rasterizer based on a deferred shading model. Per-vertex attributes, e.g., colors and normals, are computed by linear interpolation at the pixels using barycentric coordinates and triangle vertex indices. The vertex normal is calculated as the average of surrounding triangle normals [30].

The projected face albedo is first obtained using the shape, the refined texture and the pose, $\rho \in \mathbb{R}^6$, predicted by the Coarse Model. Then, the final rendered image is generated by illuminating the face mesh with the estimated lighting parameters $\gamma \in \mathbb{R}^{27}$ predicted by the CNN-based regressor in our Coarse Model. By projecting the face mesh to a 2D image, we are able to compute our loss function based on the differences between the input image and the rendered one in both image-level and perception level.

#### 3.2.4. Loss Functions

The proposed model is trained without any 3D ground truth labels. We propose a self-supervision scheme that uses losses in both image-level (photometric loss) and perception-level (identity and attribute-aware losses) in conjunction with their robustified version. The robustified losses further improve the robustness of the model to occlusions, large poses and other challenging appearance variations such as heavy make-up and shadow. It is defined as the weighted sum of the following four terms:

$$L = \alpha_{photo} L_{photo} + \alpha_{id} L_{id} + \alpha_{att} L_{att} + \alpha_{rob} L_{rob}, \quad (5)$$

where the balancing weights are empirically set to $\alpha_{photo} = 1$, $\alpha_{id} = 0.2$, $\alpha_{att} = 0.005$, and $\alpha_{rob} = 0.001$, in all our experiments.

**Photometric Loss.** As the main objective of the network is to reconstruct faces with high fidelity, we enforce the model to minimize the differences between the original input image and the rendered image obtained by the differentiable rendering layer. During the loss computation, we exclude the occluded regions, using a binary mask $M$, in order to mitigate the adverse impact of occlusions in in-the-wild images. This mask is acquired by a pre-trained face segmentation network [48]. Thus, the pixel-wise photometric loss between the input image $I$ and the rendered on $I^R$ is defined as:

$$L^{I,I^R}_{photo} = \frac{\sum M \bigodot \left\| I - I^R \right\|_2}{\sum M}, \quad (6)$$

where $\bigodot$ is the element-wise product. This loss term, which has been extensively used in the literature, is effective in supporting other loss terms with fine-grained texture. This is because it is calculated on the highest available resolution while images need to be down-sampled during the calculation of other losses as will be discussed later.

While this masking strategy is effective, it ignores the occluded areas, making their reconstruction difficult. However, the proposed self augmentation scheme introduced in Section 3.2.3 enables the recovery of face textures even under occlusions.

**Identity Loss.** Applying the pixel-level loss to measure image discrepancy can generally generate plausible results. However, it can fool the network to reconstruct faces that match closely at the pixel level but look unnatural and have smooth textures. In order to have a robust prediction of recognizable 3D faces, we propose to use a perception-level loss at the face feature level to further guide the training. By leveraging identity loss, we not only encourage the pixels of the rendered image to exactly match the pixels of the input, but we also force them to have similar identity feature representations. In this work, we employ FaceNet [1] as

our deep identity-related feature extractor. More specifically, given a pre-trained FaceNet face recognition network, we capture the deep features of the input image and the rendered image from a subset of the network layers, and define the loss in terms of the cosine distance between them, i.e.,:

$$L_{id}^{I,I^R} = 1 - \frac{F(I)F(I^R)}{\left\|F(I)\right\|_2 \left\|F(I^R)\right\|_2}. \quad (7)$$

Here, $F(.)$ denotes the deep feature encoding. Identity loss helps to learn the facial identity without being influenced by appearance factors such as facial expressions, poses, lighting and occlusions.

**Face Attribute-aware loss.** Identity features are typically extracted using face recognition networks trained to discard, throughout the convolutional layers, face attributes such as age, pose, facial expressions, and illumination. To capture some of the informative mid-level features that are useful for 3D face reconstruction, Gecer et al. [23] proposed to leverage intermediate representations by minimizing the distance of intermediate activations (extracted from a pre-trained face recognition network) of the input and the rendered image. However, it is not clear which layer exactly represents those specific attribute-related features. To address this issue, we devise a novel loss term, named face attribute-aware loss, that enforces the closeness between the input image and the reconstructed one in terms of face-related attributes. It is defined as:

$$L_{att}^{I,I^R} = \left\|F(I) - F(I^R)\right\|_2. \quad (8)$$

Here, $F(.)$ denotes the extracted attribute features.

**Robustified loss.** We further improve our perceptual loss, i.e., the identity loss and the face attribute-aware loss, using an additional perceptual loss on the rendered images, $\widehat{I}^R$, with multiple random poses for each face. Adding perceptual loss with random projections ensures that our reconstructed face resembles the target in terms of identity and attribute under various viewpoints. Also, to make the model more robust to diverse variations, e.g., large occlusions, we also use the perceptual loss on synthetically rendered images with occlusions as we realized that robust reconstruction of recognizable faces can be facilitated by applying such robustified perceptual losses. The rationale behind this is that the face identity and face attributes should be consistent across one's different images under various viewing conditions, e.g., occlusions and different poses. In summary, the robustified loss is defined as:

$$L_{rob}^{I,\widehat{I}^R} = L_{id}^{I,\widehat{I}^R} + L_{att}^{I,\widehat{I}^R}. \quad (9)$$

Here, $L_{id}^{I,\widehat{I}^R}$ and $L_{att}^{I,\widehat{I}^R}$ are identity loss and attribute-aware loss computed between the input image $I$ and the rendered image under randomly taken views or synthetic occlusions $\widehat{I}^R$.

# 4. Experimental Results

We evaluate the proposed model qualitatively and quantitatively on CelebA [49] and AFLW2000 [17] datasets and thoroughly compare our results with state-of-the-art 3D face reconstruction methods, namely MoFa [15], PRNet [16], Deep 3D Face Reconstruction [32], Nonlinear face reconstruction [21], RingNet [35], FML [37], MGCNet [26], INORig [27], GANfit [23], DF2Net [29], DECA [42], the methods of Tewari et al. [34], Egger et al. [39] and Lin et al. [14].

## 4.1. Implementation Details

We trained the proposed model using CelebA dataset [49], which contains more than 200K celebrity images. The training images were aligned following the method of [50], segmented by the pre-trained model of [48] and resized to $224 \times 224$ before being fed to the model. During the training process, we used Adam optimizer with the initial learning rate of $10^{-4}$ and a learning rate decay of 0.9, batch size of 16 and around 200K total iterations. Any 3DMM coefficient regressor could potentially be used to reconstruct a coarse 3D face (See Section 3.1). Our choice of using Deep3DDM [32] is motivated by its ability to generate a plausible coarse 3D face from unconstrained images. All experiments were performed on a PC with Intel Core i9 CPU at 3.7 GHz and a NVIDIA GeForce RTX 2080 Ti GPU.

## 4.2. Qualitative Comparison

We qualitatively evaluated and compared our method against the most recent face reconstruction methods. Performance evaluation was carried out on a subset of unseen images from CelebA dataset and AFLW2000 dataset [17]. We use the publicly available implementations of [16, 32, 21, 35, 26, 27, 29, 14, 42] to report their results and for the methods of [23, 37, 15, 34, 39], their authors provided us with the qualitative results on a subset of test set as the corresponding source codes were not publicly available.

Fig. 2 shows the superiority of our proposed method in its ability to reconstruct high-quality photo-realistic face textures even under challenging viewing conditions, e.g., extreme lighting and shadows. The third row in this figure shows the results of [14] which appears to be sensitive to viewing conditions (e.g., occlusions and lighting) and cannot extract fine textures in the presence of shadows and occlusions. This is due to the fact that it employs the exact values of the input face during the refinement step. Tran et al. [21] proposed to learn additional proxies as a means to avoid strong regularization. However, their non-linear model is still not able to reconstruct plausible faces as it was trained on 300W-LP dataset, which was generated using a linear 3DMM. The weakly supervised method of Deng et al. [32] is able to recover a basic shape and texture. However, facial details such as those in eyes and skin texture (lentigo) are not faithfully estimated. Also, in general, the reconstructed face lacks realism specifically in terms of face expressions. These results also show how this method is limited when dealing

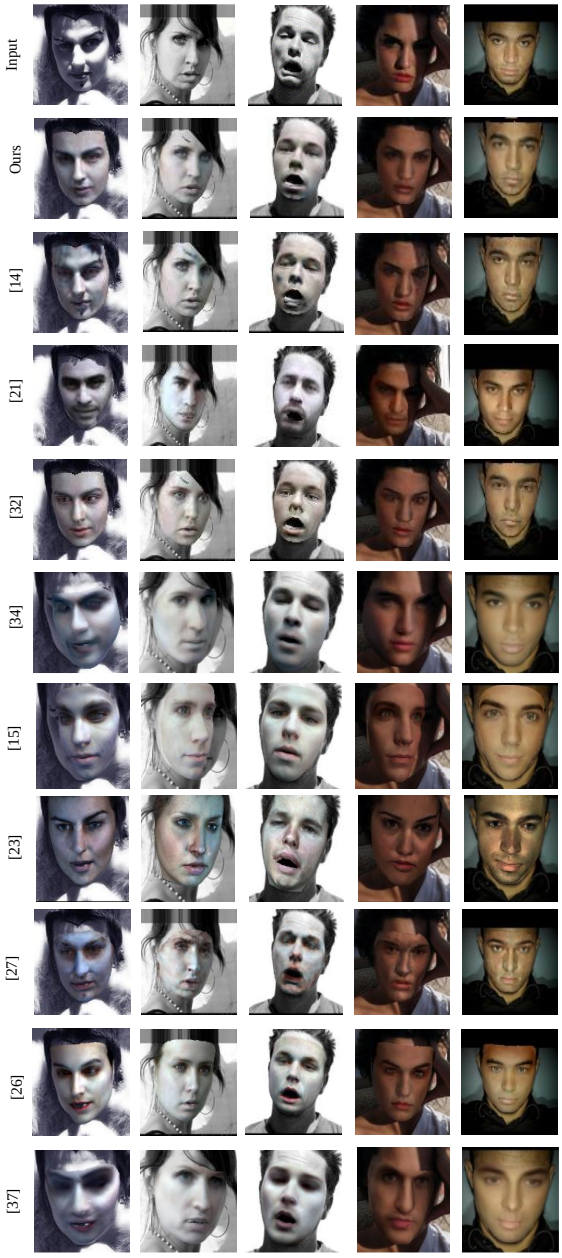


**Figure 2:** Qualitative comparison with [14, 21, 32, 34, 15, 23, 27, 26, 37] on AFLW2000 dataset. Extreme illumination conditions harm the estimation of 3D faces in some of these methods.

with in-the-wild textures, since the model is still restricted to the linear subspace.

The method of INORig [27] was developed for multi-view face reconstruction. Since only one view (image) is available for monocular face reconstruction, we horizontally flip it to generate the second input image before feeding both images to the trained model. Because the model is trained with synthetic images, it cannot generalize well when applied to in-the-wild images, resulting in texture artifacts ($9^{th}$ row). Although the monocular reconstruction works by Tewari et al. [34, 15] used end-to-end training on in-the-wild

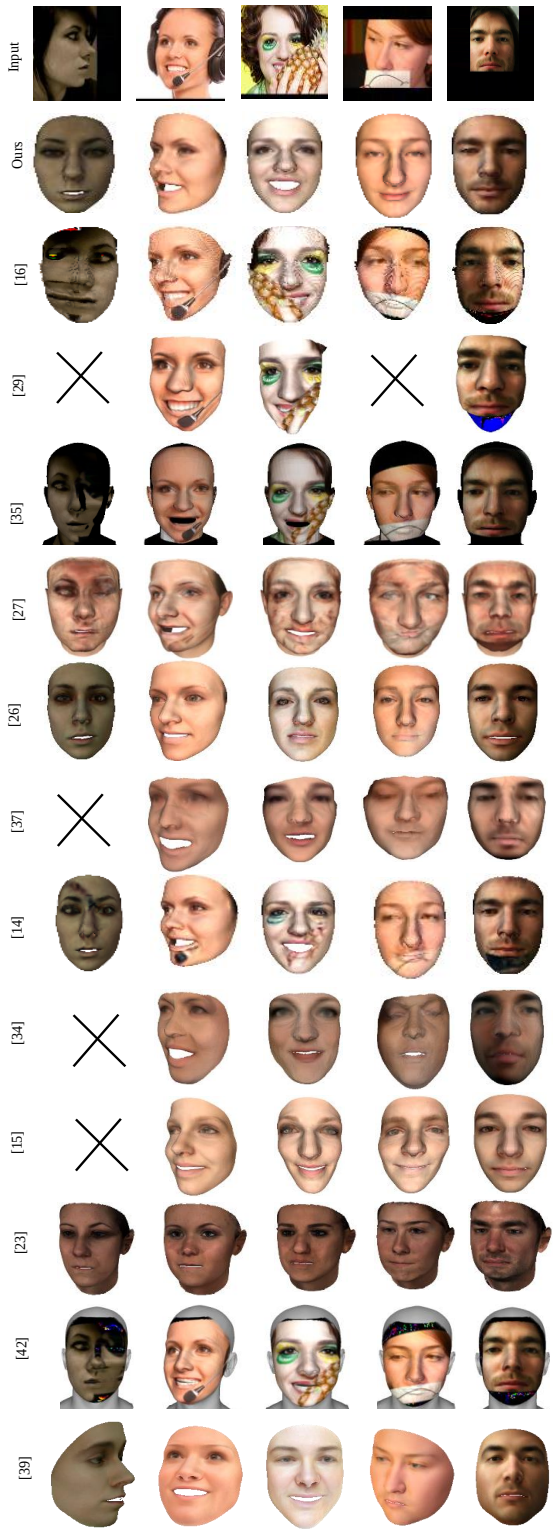


**Figure 3:** Comparison with PRNet [16], DF2Net [29], RingNet [35], INORig [27], MGCNet [26], FML [37], MOFA [15], GANfit [23], DECA [42], the methods of Lin et al. [14], Tewari et al. [34], and Egger et al. [39] on AFLW2000. Extreme head poses and severe occlusions adversely impact the reconstruction quality of these methods.

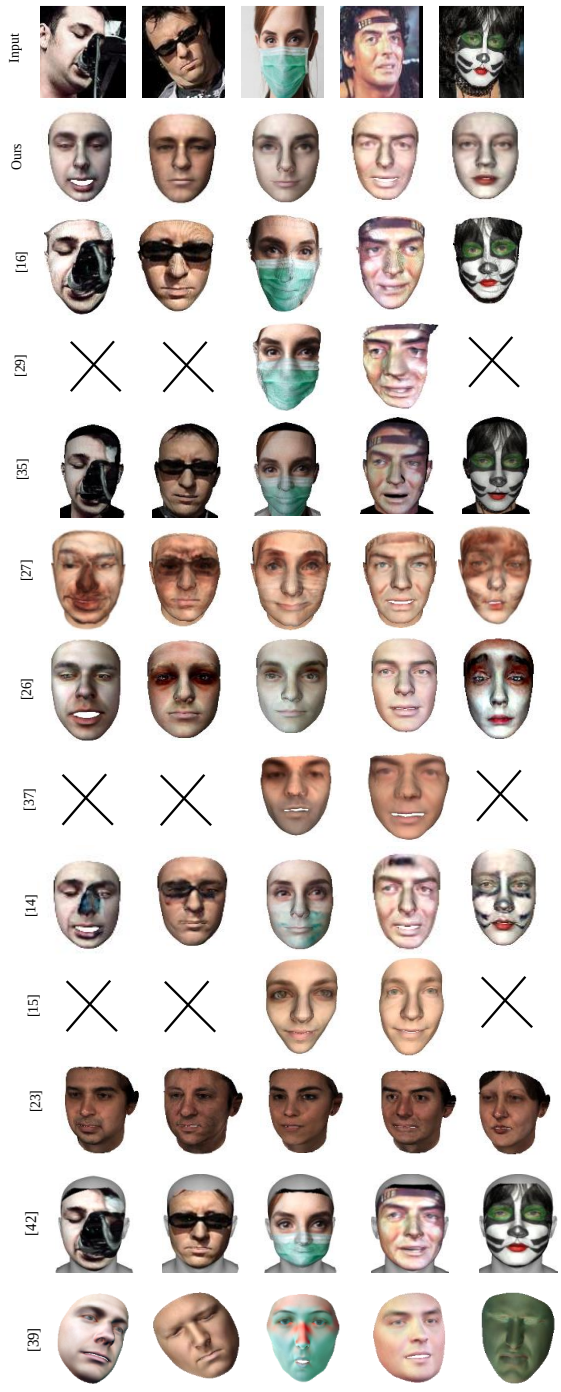


**Figure 4:** Qualitative comparison with PRNet [16], DF2Net [29], RingNet [35], INORig [27], MGCNet [26], FML [37], MOFA [15], GANfit [23], DECA [42], the methods of Lin et al. [14], and Egger et al. [39] on CelebA dataset under challenging viewing conditions.

images, it is still not able to faithfully reconstruct the identity, skin tone, and lighting in challenging viewing conditions ($6^{th}$ and $7^{th}$ rows). Its improved version [37], which uses multi-view supervision during training, also fails to capture adequate facial details and identity for monocular reconstruction. MGCNet [26] and GANFIT [23] reconstructions are also limited by the use of a pretrained 3DMM model and hence result in less detailed texture than our model which leverages fine-scale correction to add more details to the coarse–scale linear model. More specifically, the proposed method is seen to predict coloration and skin tone more faithfully. As a result, it can handle different ethnicities, genders, and scene conditions, and produce high quality reconstructions that look more realistic.

Figs. 3 and 4, show our model's ability to reconstruct fine texture details from faces under challenging appearance variations. We compared the performance of our method with recent state-of-the-art methods on AFLW2000 and CelebA datasets by reporting the reconstructed 3D meshes visualized using MeshLab. Reconstructions in the presence of extreme poses, occlusions (e.g., glasses) and heavy make-up (Figs. 3 and 4) is particularly challenging for techniques, that directly employ RGB values of the input face image to reconstruct face textures [14, 16, 29, 35]. In contrast, by extracting rich face-related features, our model is able to faithfully reconstruct all facial areas including regions hidden by strong occlusions or head rotations.

By incorporating the multi-view consistency from geometry, pixel, and depth, MGCNet [26] achieves plausible face reconstructions. However, its limitations can still be observed on reconstructed expressions (mouth shapes) and also on recovered face textures for occluded areas. PRNet [16] predicts unconstrained faces by regressing a 2D representation, i.e. UV position maps, in a supervised fashion. However, it has difficulty in reconstructing hidden regions as well as fine texture details because of the smoothing effect of the image-to-UV position mapping. DECA [42] is tolerant to common noise and occlusions exist in its training dataset and fails to recover unobserved area in extreme occlusions, head pose and lighting. The 3DMM-based method of [39] is limited by the 3DMM texture model and only captures low-frequency details. This method also has difficulties in handling extreme viewing conditions such as wearing face

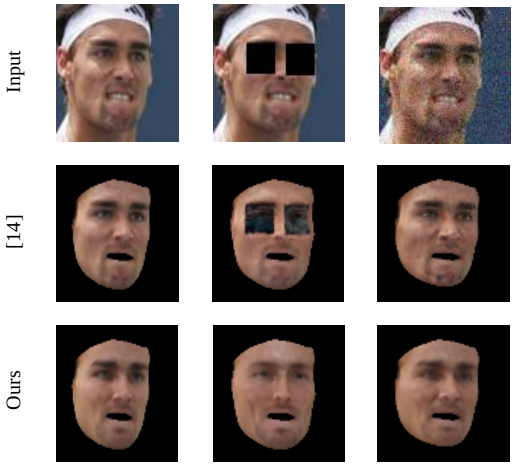


**Figure 5:** Reconstruction results from corrupted faces. Our model is robust to change in occlusion (b) and noise (c) in comparison with [14].

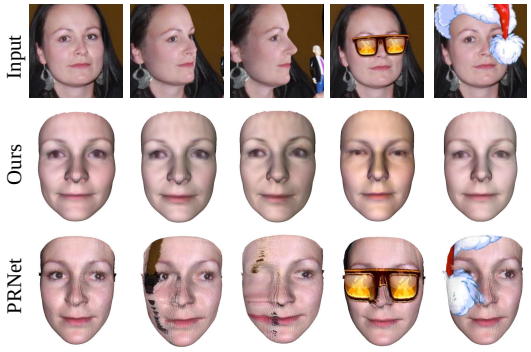


**Figure 6:** 3D face reconstruction for a single subject under different viewing conditions. Our proposed method is robust to various view-points and occlusions in comparison with PRNet [16] and is able to reconstruct all facial areas including the unobserved region.

mask (3rd column in Fig. 4) and heavy make-up (5th column in Fig. 4). As shown in Figs. 3 and 4, our method outperforms all other methods even in the presence of large poses, (self-) occlusions and heavy make-up.

We further evaluate qualitatively the robustness of the proposed method to diverse variations by conducting an experiment that reconstructs 3D faces from synthetically corrupted images (See Fig. 5). To better understand the superiority of our method, we compare our results with those of Lin et al. [14]. As shown in Fig. 5, [14] cannot recover plausible face textures from faces with challenging viewing conditions. However, our model, which employs a rich face representation describing face attributes and identity as the input for the texture refiner, is robust to noise and extreme occlusions. In addition, Fig. 6, qualitatively demonstrates this robustness in compariosn with [16] by varying extreme conditions (various viewpoints and occlusions) for a single subject and displaying consistent outputs for our model. This robustness is mainly due to the invariance of the extracted high-level discriminative features to pixel-level information.

We also evaluate our approach on a set of images with challenging expressions, which often cause wrinkles and folds in the face; see Fig. 7. As can be seen in this figure, by applying the Fine Model to the initial coarse 3D face, our model is able to nicely recover the subtle texture details such as small crow's feet, wrinkles, and folds.

### 4.3. Quantitative Comparison

To conduct a quantitative analysis, we compute the metrics on rendered 2D images due to the lack of 3D ground truth data. Specifically, we calculate Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), which measures the similarity between an input image and the projected image from 3D face mesh. To better understand the power of our model to reconstruct with high fidelity faces in challenging viewing conditions, e.g., in the presence of occlusions, we evaluate our method on synthetically occluded images of CelebA and AFLW2000 in which the non-occluded face images are available as the reference.

**Table 1**
Quantitative comparison results on CelebA and AFLW2000. The results of [32], [14], [16], [26] and [27] are obtained using the original implementations provided by the authors. Best results in bold.

| Dataset | Method | Quality Metrics | | | | ID Similarity | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR | | SSIM | | FaceNet | | LightCNN | | EvoLVe | |
| | | Mean | Std. | Mean | Std. | Mean | Std. | Mean | Std. | Mean | Std. |
| CelebA | [32] | 22.47 | **2.239** | 0.826 | 0.035 | 0.890 | 0.034 | 0.51 | 0.146 | 0.32 | **0.14** |
| | [14] | 18.59 | 2.582 | 0.817 | 0.035 | **0.953** | 0.026 | 0.48 | 0.174 | 0.39 | 0.19 |
| | [16] | 17.96 | 2.51 | 0.802 | 0.037 | 0.903 | 0.031 | 0.515 | 0.14 | 0.35 | 0.17 |
| | [26] | 18.86 | 2.47 | 0.79 | 0.041 | 0.893 | 0.029 | 0.49 | 0.154 | 0.35 | 0.149 |
| | [27] | 17.88 | 2.45 | 0.77 | 0.037 | 0.88 | 0.034 | 0.508 | 0.133 | 0.341 | 0.178 |
| | Ours | **23.27** | 2.38 | **0.846** | **0.034** | 0.938 | **0.017** | **0.56** | **0.126** | **0.40** | **0.14** |
| AFLW2K | [32] | 23.21 | 2.769 | 0.856 | 0.034 | 0.908 | 0.029 | 0.63 | 0.130 | 0.42 | 0.15 |
| | [14] | 19.02 | 3.147 | 0.844 | 0.025 | **0.955** | **0.020** | 0.60 | 0.159 | 0.44 | 0.18 |
| | [16] | 18.08 | 2.80 | 0.811 | 0.025 | 0.913 | 0.025 | 0.61 | 0.151 | 0.32 | 0.15 |
| | [26] | 18.36 | 2.84 | 0.809 | 0.033 | 0.941 | 0.024 | 0.607 | 0.134 | 0.38 | 0.17 |
| | [27] | 17.74 | 2.65 | 0.79 | 0.029 | 0.901 | 0.028 | 0.61 | 0.158 | 0.38 | 0.18 |
| | Ours | **23.84** | **2.581** | **0.871** | **0.024** | 0.933 | 0.022 | **0.65** | **0.124** | **0.44** | **0.14** |

Higher PSNR and SSIM reported in Table 1 indicate that our reconstructed faces are closer to the input images in pixel-level and are more consistent from person to person with less standard deviation compared to other methods.

In order to evaluate the recognizability of the faces reconstructed with our approach, we employ a set of face recognition features as a measure of similarity. In particular, for each face image, we compute the cosine similarity of the layers of FaceNet between the input image and the rendered one. In order to avoid bias in evaluating the perceptual likeness, two other pre-trained face recognition networks, i.e., LightCNN [51] and EvoLVe [52], have been employed to extract discriminative perceptual face representations. The results in Table 1 indicate that our model has the most identity-preserving ability, thanks to applying the robustified perception-level loss.


**Figure 7:** Illustration of the model's performance under various expressions by overlaying the reconstructions on the input images. Our method can generate high fidelity reconstructions under large variations of expression.

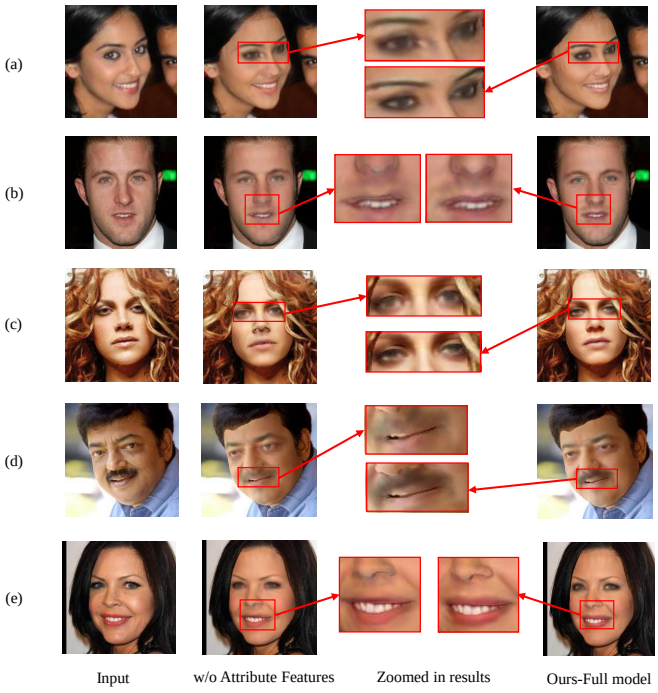


**Figure 8:** Illustration of the effects of employing Face Attribute Features in our Fine Model. The reconstructed textures in our full model are very good at capturing high frequency details in eyes, noses and lips.

## 4.4. Ablation Study

We perform an ablation study on CelebA dataset by evaluating several variants of our model. To explore how the performance is affected by each factor, we train our model with various combinations of components or loss terms and report the obtained qualitative and quantitative results.

Fig. 8 demonstrates the effect of leveraging face attribute features in our Fine Model. We found that employing such features helps to reduce artifacts by providing smoother surfaces but still allowing high frequency details to be preserved. The reconstructed textures in our full model are very good at capturing high frequency details in eyes, noses and lips (Fig. 8 (a), (b) and (e)), with sharp edges. Removing this set of features results in discontinuities and artifacts in those parts. Moreover, the attribute preserving power of our full model can be seen in Fig. 8 (c) and (d) in which the languishing eyes and mustaches are faithfully reconstructed and look more realistic thanks to employing a set of discriminative attribute-aware features in our Feature-wise Refiner.

We also demonstrate the influence of our self-augmentation technique by switching off the self-augmentation module and evaluating qualitatively the performance of the ablated model. Fig. 9(a) shows visual comparisons between our full model and its variant, i.e., ours without applying self-augmentation technique in the Fine Model. To better understand the effect of this module in generating plausible estimates for the parts that are completely hidden from camera, we specifically demonstrate the reconstructed 3D meshes in unobserved regions of faces viewed under out-of-plane rotation. It can be seen from Fig. 9(a) that our self-augmentation module helps to generate plausible and naturally looking facial details for unobserved facial regions and leads to improved reconstruction quality. More specifically, Fig. 9(b) shows the effect of using synthetically occluded rendered images in our self-augmentation module. We observe that the ablated model fails to recover the regions hidden by occlusions, for example, the hair-covered forehead and eyebrows.

**Table 2**
Contributions of the components of the proposed approach on CelebA dataset (Higher is better). SA refers to Self-Augmentation and FA refers to Face Attribute.

| | PSNR | SSIM | ID Similarity | | |
|---|---|---|---|---|---|
| | | | FaceNet | LightCNN | EvoLVe |
| $L_{photo} + L_{id}$ | 22.78 | 0.831 | 0.918 | 0.539 | 0.3685 |
| $L_{photo} + L_{id} + L_{att}$ | 22.91 | 0.837 | 0.9221 | 0.550 | 0.376 |
| $L_{photo} + L_{id} + L_{att} + L_{rob}$ | **23.27** | **0.846** | **0.938** | **0.56** | **0.3956** |
| Full model w/o SA module | 21.19 | 0.805 | 0.912 | 0.55 | 0.374 |
| Full model w/o FA features | 20.14 | 0.810 | 0.925 | 0.517 | 0.380 |

To explore how individual terms of our loss function affect the overall performance, we start training our model by the photometric loss and identity loss and then add other terms one by one and compute the aforementioned metrics for each variant. Table 2 reports the quantitative results for each variant. One can note that each of the loss function terms contributes towards an improved reconstruction quality. As expected, our novel loss terms, i.e., face attribute-aware loss, $L_{att}$ and robustified loss, and $L_{rob}$, are effective in improving the performance. However, $L_{rob}$ appears to have more contribution than the others in boosting the performance especially in terms of system's recognizability as measured by identity similarity. We also quantitatively

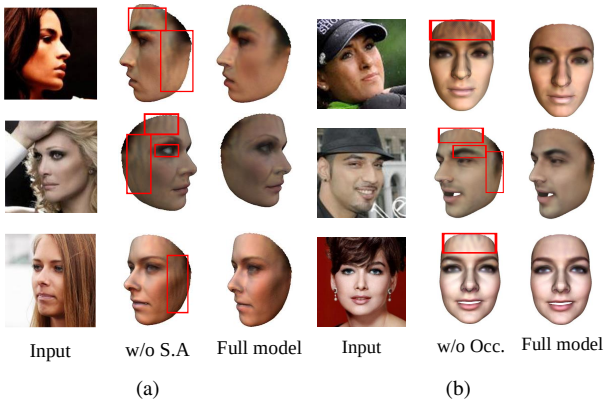


**Figure 9:** Illustration of the benefits of applying Self-augmentation technique in our Fine Model. Our full model obtains higher reconstruction quality and provides robustness to occlusions and strong head rotation. SA refers to Self-Augmentation and Occ. refers to Occlusion.

evaluated the importance of the Self Augmentation (SA) module as well as that of employing Face Attribute (FA) features in our model. The obtained results (reported in Table 2) indicate that the highest similarity (in terms of PSNR, SSIM and ID similarities) is achieved by the full model.

## 5. Conclusions and Future Work

In this paper, we propose a fully unsupervised method for 3D face reconstruction. Delicate facial textures, which are difficult to represent by a 3DMM, are reconstructed by applying a novel Fine Model to the initial coarse 3D face. To supervise and facilitate the training, we employ a set of effective self-supervised loss functions for efficient end-to-end training of the architecture on a large corpus of in-the-wild face images. The qualitative and quantitative experiments on two challenging face datasets indicate that, compared to existing methods, the proposed approach is robust to viewing conditions and is able to reconstruct higher fidelity 3D faces without the need for any annotated ground truth. We have also tackled diverse appearance variations, e.g., wearing heavy make-up or glasses. As shown in the qualitative results, our model has the ablity to successfully remove the make-up, which makes it ideal for biometric applications that require make-up-invariant face verification. However, in some applications such as Augmented/Virtual Reality (AR/VR), to have a much more immersive AR/VR experience, or first-person experience in general, we need to reconstruct the person's face with its current appearance. Therefore, completely removing the make-up or glasses in such applications would deteriorate the immersive experience. Future work should focus on recognizing such variations and trying to reconstruct them on top of the initially reconstructed face. This suggests the need for an intelligent approach to properly model such variations.

## Acknowledgement

This work is supported by ARC DP210101682. We thank authors who made their source codes publicly available or provided us with their results and made the comprehensive comparison possible.

## References

[1] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[2] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2018.

[3] Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling features in 3d face shapes for joint face reconstruction and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5216–5225, 2018.

[4] Jian Zhao, Lin Xiong, Jianshu Li, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 3d-aided dual-agent gans for unconstrained face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2380–2394, 2018.

[5] Jie Liang, Huan Tu, Feng Liu, Qijun Zhao, and Anil K Jain. 3d face reconstruction from mugshots: Application to arbitrary view face recognition. *Neurocomputing*, 410:12–27, 2020.

[6] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[7] Roger Blanco i Ribera, Eduard Zell, John P Lewis, Junyong Noh, and Mario Botsch. Facial retargeting with automatic range of motion alignment. *ACM Transactions on graphics (TOG)*, 36(4):1–12, 2017.

[8] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. *ACM Transactions on Graphics (ToG)*, 32(4):1–10, 2013.

[9] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 fourth international conference on 3D vision (3DV)*, pages 460–469. IEEE, 2016.

[10] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1259–1268, 2017.

[11] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5908–5917, 2017.

[12] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE international conference on computer vision*, pages 1031–1039, 2017.

[13] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017.

[14] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2020.

[15] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017.

[16] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018.

[17] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[18] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016.

[19] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library, 2018.

[20] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[21] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019.

[22] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017.

[23] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019.

[24] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *arXiv preprint arXiv:2105.07474*, 2021.

[25] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7093–7102, 2018.

[26] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 53–70. Springer, 2020.

[27] Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, and Ping Tan. Riggable 3d face reconstruction via in-network optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6216–6225, 2021.

[28] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018.

[29] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2315–2324, 2019.

[30] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018.

[31] Xiaoguang Tu, Jian Zhao, Mei Xie, Zihang Jiang, Akshaya Balamurugan, Yao Luo, Yang Zhao, Lingxiao He, Zheng Ma, and Jiashi Feng. 3d face reconstruction from a single image assisted by 2d face images in the wild. *IEEE Transactions on Multimedia*, 2020.

[32] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[33] Anil Bas, Patrik Huber, William AP Smith, Muhammad Awais, and Josef Kittler. 3d morphable models as spatial transformer networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 904–912, 2017.

[34] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018.

[35] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019.

[36] Jian Zhang, Ke Li, Yun Liang, and Na Li. Learning 3d faces from 2d images via stacked contractive autoencoder. *Neurocomputing*, 257: 67–78, 2017.

[37] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019.

[38] Bernhard Egger, Andreas Schneider, Clemens Blumer, Andreas Forster, Sandro Schönborn, and Thomas Vetter. Occlusion-aware 3d morphable face models. In *BMVC*, volume 2, page 4, 2016.

[39] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018.

[40] Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. To fit or not to fit: Model-based face reconstruction and occlusion segmentation from weak supervision. *arXiv preprint arXiv:2106.09614*, 2021.

[41] Hitika Tiwari, Min-Hung Chen, Yi-Min Tsai, Hsien-Kai Kuo, Hung-Jen Chen, Kevin Jou, KS Venkatesh, and Yong-Sheng Chen. Self-supervised robustifying guidance for monocular 3d face reconstruction. *arXiv preprint arXiv:2112.14382*, 2021.

[42] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[43] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3): 413–425, 2013.

[44] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *arXiv preprint arXiv:1606.09375*, 2016.

[45] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. *arXiv preprint arXiv:1808.06882*, 2018.

[46] Baris Gecer, Binod Bhattarai, Josef Kittler, and Tae-Kyun Kim. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–234, 2018.

[47] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3703–3712, 2017.

[48] Tianyang Shi, Yi Yuan, Changjie Fan, Zhengxia Zou, Zhenwei Shi, and Yong Liu. Face-to-parameter translation for game character auto-creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 161–170, 2019.

[49] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[50] Dong Chen, Gang Hua, Fang Wen, and Jian Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision*, pages 122–138. Springer, 2016.

[51] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[52] Jian Zhao, Jianshu Li, Xiaoguang Tu, Fang Zhao, Yuan Xin, Junliang Xing, Hengzhu Liu, Shuicheng Yan, and Jiashi Feng. Multi-prototype networks for unconstrained set-based face recognition. *arXiv preprint arXiv:1902.04755*, 2019.

Hoda Mohaghegh received the B.Sc. and M.Sc. degrees in Electrical Engineering from Isfahan University of Technology, Isfahan, Iran, in 2012 and 2016, respectively. She is currently a PhD candidate in the School of Computer Science and Software Engineering, The University of Western Australia. Her main research interests are in the field of 2D and 3D image processing, computer vision and machine learning with an emphasis on

monocular depth estimation and 3D reconstruction.

Faird Boussaid received the M.S. and Ph.D. degrees in microelectronics from the National Institute of Applied Science (INSA), Toulouse, France, in 1996 and 1999 respectively. He joined Edith Cowan University, Perth, Australia, as a Postdoctoral Research Fellow, and a Member of the Visual Information Processing Research Group in 2000. He joined the University of Western Australia, Crawley, Australia, in 2005, where he is currently a Professor. His current research interests include neuromorphic engineering, smart sensors, and machine learning.

Hamid Laga received the PhD degrees in Computer Science from Tokyo Institute of Technology in 2006. He is currently a Professor at Murdoch University (Australia). His research interests span various fields of machine learning, computer vision, computer graphics, and pattern recognition, with a special focus on the 3D reconstruction, modeling and analysis of static and deformable 3D objects, and on image analysis and big data in agriculture and health. He is the recipient of the Best Paper Awards at SGP2017, DICTA2012, and SMI2006

Hossein Rahmani received the BSc degree in computer software engineering from Isfahan University of Technology, Isfahan, Iran, in 2004, the MSc degree in software engineering from Shahid Beheshti University, Tehran, Iran in 2010, and the PhD degree from the University of Western Australia, in 2016. He has published several papers in top conferences and journals such as CVPR, ICCV, ECCV, and the IEEE Transactions on Pattern Analysis and Machine Intelligence. He is currently an associate professor (Senior Lecturer) in the School of Computing and Communications at Lancaster University. Before that he was a research fellow in the School of Computer Science and Software Engineering, University of Western Australia. His research interests include computer vision, action recognition, 3D shape analysis, and machine learning.

Mohammed Bennamoun is a Winthrop Professor in the Department of Computer Science and Software Engineering at the University of Western Australia (UWA) and is a researcher in computer vision, machine/deep learning, robotics, and signal/speech processing. He has published 4 books (available on Amazon), 1 edited book, 1 Encyclopedia article, 14 book chapters, 180+ journal papers, 260+ conference publications, 16 invited and keynote publications. His h-index is 69 and his number of citations is 21,470+ (Google Scholar). He was awarded 70+ competitive research grants, from the Australian Research Council, and numerous other Government, UWA, and industry Research Grants. He successfully supervised 35+ Ph.D. students to completion. He won the Best Supervisor of the Year Award at Queensland University of Technology (1998) and received the award for research supervision at UWA (2008 and 2016) and Vice-Chancellor Award for mentorship (2016). He delivered conference tutorials at major conferences, including IEEE CVPR 2016, Interspeech 2014, IEEE ICASSP, and ECCV. He was also invited to give a Tutorial at an International Summer School on Deep Learning (DeepLearn 2017).