Occlusion Resistant Network for 3D Face Reconstruction

Hitika Tiwari *† hitika@iitk.ac.in
Vinod K. Kurmi‡ vinod.kurmi@kuleuven.be
K.S. Venkatesh † venkats@iitk.ac.in
Yong-Sheng Chen* yschen@cs.nycu.edu.tw

Abstract

3D face reconstruction from a monocular face image is a mathematically ill-posed problem. Recently, we observed a surge of interest in deep learning-based approaches to address the issue. These methods possess extreme sensitivity towards occlusions. Thus, in this paper, we present a novel context-learning-based distillation approach to tackle the occlusions in the face images. Our training pipeline focuses on distilling the knowledge from a pre-trained occlusion-sensitive deep network. The proposed model learns the context of the target occluded face image. Hence our approach uses a weak model (unsuitable for occluded face images) to train a highly robust network towards partially and fully-occluded face images. We obtain a landmark accuracy of 0.77 *against* 5.84 *of recent state-of-the-art-method for real-life challenging facial occlusions. Also, we propose a novel end-to-end training pipeline to reconstruct 3D faces from multiple variations of the target image per identity to emphasize the significance of visible facial features during learning. For this purpose, we leverage a novel composite multi-occlusion loss function. Our multi-occlusion per identity model shows a dip in the landmark error by a large margin of* 6.67 *in comparison to a recent state-of-the-art method. We deploy the occluded variations of the CelebA validation dataset and AFLW2000-3D face dataset: naturally-occluded and artificially occluded, for the comparisons. We comprehensively compare our results with the other approaches concerning the accuracy of the reconstructed 3D face mesh for occluded face images.*

1. Introduction

3D face reconstruction has been a longstanding problem in the field of computer vision and computer graphics. However, in this work, we focus on addressing the problem of occlusions in monocular face images. The problem of 3D face reconstruction from occluded face images has numerous applications viz, dimension reduction of the search space for the face recognition models and interactions in the virtual environments.

In the literature, one category of approaches [24, 4, 47, 20, 2] matches the 3D face with the monocular face image by optimizing pose, texture, shape, and illumination coefficients. Another category exploits Convolution Neural Network (CNN) and Generative Adversarial Network (GAN) [17] based approaches [27, 19, 30, 35, 18, 10, 16, 36, 39, 40, 42, 44, 8, 26, 9, 15, 25, 38] rather than the optimization-based method due to the implicated complexities of optimization approach. These methods perform poorly with the artificially and naturally occluded face images and monocular face images with unusual lighting. Further, they adapt occlusions as facial features thus lead to *ghost-like* 3D face reconstruction from occluded face images.

We address the discussed issues using two training frameworks. We refer to the first model as the **Single-Occlusion per Identity** model. This model exploits one occluded image per person to learn the deep 3D face representation. The optimization and deep network-based regression methods depend upon learning 3D faces via facial features in 2D images. However, in our case, the input monocular face images to the model are completely (or partially) occluded. Thus, the proposed framework learns from the global context and local context of the images [45]. We distill the knowledge [21] of RNet [10] for unoccluded face images to train our model on the occluded images. Hence, the proposed Single-Occlusion per Identity method relies upon context-based distillation learning during the training.

Further, we propose a novel method to predict 3D face mesh from the multiple variations with regards to shape, color, and spatial location of the occlusions of face image per person. We refer to the framework as the **Multi-Occlusion per Identity** model. The model enables higher accuracy of the reconstructed 3D face mesh from occluded face images.

We evaluate our approaches on the validation dataset of CelebA [28] (released by Tensorflow) and AFLW2000-3D dataset [46]. For this purpose, we occlude the face images in the datasets. We compare the proposed models with other state-of-the-art methods. By exploiting the **Single-Occlusion per Identity** model, our approach outperforms

*National Yang Ming Chiao Tung University, Taiwan
†Indian Institute of Technology Kanpur, India
‡KU Leuven, Belgium

the current state-of-the-art methods by large margins. We deploy the CelebA validation dataset [28] and AFLW-2000 3D dataset [46] for the evaluation of the proposed model. For the naturally occluded face images (real-life facial occlusions), the landmark inaccuracy falls to $\mathbf{0.77}$, as against to $\mathbf{5.84}$ of a recent state-of-the-art method. Further, the proposed **Multi-Occlusion per Identity** model significantly improves the network performance for occluded face images. The landmark error for our multi-occlusion per identity model reduces by a factor of $\mathbf{5.59}$ as compared to a recent state-of-the-art method on the CelebA validation dataset.

We summarize our contributions in this paper below.

1. **Single-Occlusion per Identity Network for Occlusion Robust 3D Face Reconstruction**: We reconstruct the 3D faces from the *heavily occluded* 2D face images. Our model exploits RNet from [10] to predict 3D faces from a purely or partially occluded 2D face images. Thus, facilitating the learning of the proposed model through context in the target images.

2. **Multi-Occlusions per Identity Network for Improved Occlusion Robust 3D Face Reconstruction Accuracy**: We propose a method to learn 3D faces from multiple artificially occluded face images per person to improve the accuracy of the reconstructed 3D face meshes from the occluded images. Our model learns a dense representation of 3D faces using a novel composite multi-occlusion loss.

3. **Artificial Occlusions based Dataset**: We synthesize a large-scale artificial occlusion-based dataset from unoccluded and naturally occluded in-the-wild face images. The artificial occlusion-based dataset facilitates the occlusion robust training of the proposed models. We refer to goggles, facial hair, noise, etc., as natural (real-life) occlusions.

2. Related Work

3D face reconstruction from a monocular face image is a challenging research problem. Blanz and Vetter in [5] introduce a morphable model, referred to as the 3DMM, to reconstruct 3D face mesh from the monocular face images. The 3DMM plays a crucial role in addressing the problem. In 3DMM, a set of parameters accredits the reconstruction of a 3D face mesh. One can acquire the set of parameters accrediting a 3D face mesh using two approaches: optimization-based and deep network-based.

Optimization-based Methods: The approaches in [3, 5, 13] utilize optimization-based methods to fit the 3DMM model to an image, collection of images, or video data [14, 31, 32] by iteratively optimizing the alignment between the projected 3D face and the input face image. These methods adapt mild occlusions (such as hair) as facial features thus, leading to wrong results. The performance of these methods further drop for strong occlusions such as dense beard.

Deep Network-based Methods: The methods by Chen et al. [8], Deng et al. [10, 9], Gecer et al. [15], Genova et al. [16], Guo et al. [18, 19], Lin et al. [25, 26], Liu et al. [27], Richardson et al. [30], Sela et al. [35], Sengupta et al. [36], Tewari et al. [40, 39, 38], Tran et al. [42] and Zhang et al. [44] use CNN and GAN-based regression methods due to the intricacies with the optimization-based approach. Among these, the supervised approaches [12, 18, 19, 27, 30, 35, 37] suffer from the lack of training data. Numerous CNN-based training methods [1, 11, 30, 40, 43, 16] emphasize on directly using the predicted coefficients of 3DMM, while others [30, 39, 41, 19] add corrections (or details) on coarse 3DMM predictions or focus on model-free reconstruction [12, 35, 36, 42].

Lin et al. [26, 25], Deng et al. [9, 10], Gecer et al. [15], Chen et al. [8], Tewari et al. [38, 39, 40], and Zhang et al. [44] address the problem of insufficiency of the training data. These methods rely on weak labels for tackling the inconsistency between the input image and the projected 3D face to train their respective GAN or CNN-based models. Further, the mentioned weakly supervised and (or) self-supervised approaches depend upon VGGFace2 [7] and FaceNet [34] to reconstruct a perceptually precise 3D face mesh. These models are heavily sensitive towards image occlusions. Thus leading to non-human like faces for heavily occluded target face images.

In the proposed work, we do not exploit any formerly discussed methods for solving the problem. Instead, we use a context-learning-based distillation approach using the pre-trained occlusion sensitive RNet model in [10] for training the Single-Occlusion per Identity model. We also present a novel training pipeline facilitating model learning with multiple occluded (various artificial occlusion patterns) images per person. For the training, we use the novel composite multi-occlusion loss. We refer to the network as the Multi-Occlusion per Identity model. The latter model ensures higher accuracy of the reconstructed 3D face meshes than the former. On the contrary, the latter puts constraints over the data requirement. It requires multiple occluded face images per identity for the training.

Our findings in this work are multitudinous: $\mathbf{1}$) the context-based distillation approach opens a new dimension of research for improving the model performance (in terms of the robustness towards occlusion) with simple *twists and tricks*, $\mathbf{2}$) distribution-based loss functions are unsuitable for the training the model under discussion, and $\mathbf{3}$) the multiple occlusion patterns per identity for training the model ensures higher accuracy of the reconstructed 3D faces from the occluded face images.

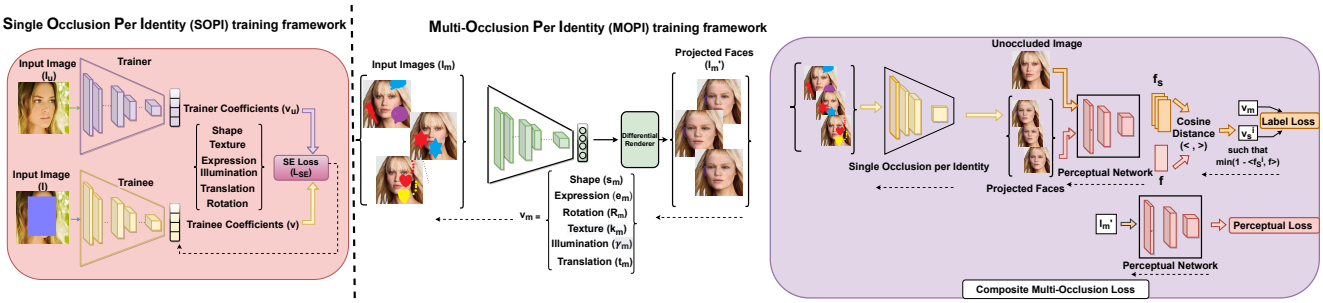


Figure 1: **a)** An overview of the training frameworks of the proposed occlusion robust 3D face reconstruction models. We deploy context-learning-based distillation approach to train **S**ingle-**O**cclusion **P**er **I**dentity (**SOPI**) Model (left). The trainer gets an unoccluded face image at the input, whereas we feed the trainee with partial/fully/unoccluded face data on a random basis. **b)** Besides, we also train a novel **M**ulti-**O**cclusion **P**er **I**dentity (**MOPI**) Model (right). The model takes multiple images of the same identity. Each image is occluded with a random occlusion pattern.

3. Proposed Model Description

In this section, we discuss the proposed training architectures. We refer to our training frameworks as **S**ingle-**O**cclusion **P**er **I**dentity (**SOPI**) (refer Sec. 3.1) and **M**ulti-**O**cclusion **P**er **I**dentity (**MOPI**) (refer Sec. 3.2) depending upon the input data. We also present the losses exploited to train the models under discussion.

3.1. Single-Occlusion per Identity Network

In Fig. 1a, we present our training model. The details of our network are discussed below.

3.1.1 Trainee Network

Our trainee framework (see Fig. 1a) uses the ResNet-50 model as backbone architecture. It takes input image $\mathbf{I}$ (partially (or fully) occluded or unoccluded) and returns the coefficients of 3DMM. We use popular Base Face Model (BFM) [29] as 3DMM. We exploit RNet from [10] to train our model. Further, we modify the fully-connected classification layer of ResNet-50 with 257 nodes for obtaining the coefficients denoted by $\mathbf{v} \in \mathbb{R}^{257}$.

$$\mathbf{v}(\theta) = \text{Single-Occlusion per Identity}(\mathbf{I}, \theta). \quad (1)$$

In Eq. (1), θ represents network weights.

3.1.2 Trainer Network

We use pre-trained RNet occlusion-sensitive model from [10] as the *trainer* network (see Fig. 1a). We freeze the weights of the RNet model during the training of the trainee model. RNet utilizes the ResNet-50 model as a backbone with the fully-connected layer of 257 nodes. The trainer model takes an unoccluded input image $\mathbf{I_u}$ and returns the coefficients ($\mathbf{v_u}$) of 3D morphable BFM. We use these labels to regress the trainee model over the coefficients of BFM. Eq. (2) depicts the input-output relationship of RNet model.

$$\mathbf{v_u} = \text{RNet}(\mathbf{I_u}). \quad (2)$$

RNet depends upon weak labels such as skin attention mask and 68 landmark coordinates of a face image for its training. It also requires a mesh-renderer, landmark detector, and the basis from BFM to perform the training of the model. RNet takes $\mathbf{I_u}$ to produce vector $\mathbf{v_u} \in \mathbb{R}^{257}$ containing shape ($\mathbf{s}$), expression ($\mathbf{e}$), texture ($\mathbf{k}$), illumination (γ) and pose ($\mathbf{R}, \mathbf{t}$) coefficients. The coefficients thus obtained serves as the input to the mesh-renderer after undergoing through mathematical analysis (see Eq. (3) for shape, expression and texture, Eq. (4) for illumination and Eq. (5) for pose).

$$\mathbf{S} = \overline{\mathbf{S}} + \mathbf{s}\mathbf{B_s} + \mathbf{e}\mathbf{B_e}, \qquad \mathbf{K} = \overline{\mathbf{K}} + \mathbf{k}\mathbf{B_k}. \quad (3)$$

In Eq. (3), $\overline{\mathbf{S}} \in \mathbb{R}^{3N}$ and $\overline{\mathbf{K}} \in \mathbb{R}^{3N}$ indicate mean face shape and skin reflectance (or texture) whereas, $\mathbf{B_s}$ and $\mathbf{B_k}$ are the subsets of Principal Component Analysis (PCA) basis for shape and texture respectively from the BFM. The expression basis $\mathbf{B_e}$ are acquired from the Facewarehouse model [6]. Further, RNet exploits the shape coefficient vector $\mathbf{s} \in \mathbb{R}^{80}$, expression vector $\mathbf{e} \in \mathbb{R}^{64}$ and texture coefficient vector $\mathbf{k} \in \mathbb{R}^{80}$ for the reconstruction of 3D face mesh.

RNet assumes a *Lambertian* surface reflectance for 3D face mesh, and represent the illumination using Spherical Harmonics (SH). Eq. (4) establishes the relationship between radiosity ($\mathbf{C}$) of vertex $\mathbf{x_i}$ with normal $\mathbf{n_i}$ and skin texture $\mathbf{K_i}$.

$$\mathbf{C}(\mathbf{K_i}, \mathbf{n_i}, \gamma) = \mathbf{K_i} \cdot \sum_{b=1}^{B^2} \gamma_\mathbf{b} \psi_\mathbf{b}(\mathbf{n_i}). \quad (4)$$

In Eq. (4), the coefficient $\gamma_\mathbf{b} \in \mathbb{R}^3$ ($B = 3$ bands) represents the red, green, and blue channels for illuminating

the 3D face mesh corresponding to SH basis coefficients $\psi_{\mathbf{b}} : \mathbb{R}^3 \rightarrow \mathbb{R}$. We assume white lights thus, $\gamma \in \mathbb{R}^9$.
To project a 3D face onto a 2D image, RNet uses a pinhole camera under full perspective projection (using mesh-renderer). Eq. (5) describes the pose of a 3D face mesh parameterized by rotation matrix $\mathbf{R} \in SO(3)$ and translation vector $\mathbf{t} \in \mathbb{R}^3$.

$$\mathbf{I_u}' = \zeta(\mathbf{R}^{-1}(\mathbf{x} - \mathbf{t})). \tag{5}$$

In Eq. (5), $\mathbf{I_u}'$ represents the image formed by projecting a 3D face mesh, $\zeta : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ maps world coordinates to the screen, and $\mathbf{x}$ is the collection of coordinates for mesh vertices.
RNet relies upon reducing the discrepancies between $\mathbf{I_u}$ and $\mathbf{I_u}'$ using image-level losses such as photometric (using skin attention mask) and landmark losses, and feature-level losses.

3.1.3 Learning Context from the Trainer

The trainer produces coefficients of 3DMM for a given unoccluded face image while forcing the *trainee* to reproduce the same coefficients for the same input face image with occlusion. To learn the 3D face image of a given input 2D face, the *trainee* has to be consistent with the trainer irrespective of the occlusion in the input face image. The approach under discussion facilitates the *trainee* to improve robustness using the knowledge distillation for occluded input. The consistency with the *trainer* criteria enforces the *trainee* model to learn the context from occluded and challenging input. It also facilitates trainee to enhance its self-learning capability. We use SE loss to minimize the difference between the predicted coefficients of *trainer* and the *trainee* models.

3.1.4 Squared Error Loss

The SE loss function enforces the trainee coefficients to be close to the trainer coefficients for an image. We provide an occluded (or unoccluded) image as the input to our trainee model. Further, we compare these coefficients with the labels for the corresponding unoccluded image acquired from the trainer model. Eq. (6) depicts the SE loss L_{SE}.

$$\begin{aligned} \mathbf{c} &= (\mathbf{v_u} - \mathbf{v})^2, \\ L_{SE} &= \frac{1}{M} \sum_{m=1}^{M} c_m. \end{aligned} \tag{6}$$

In Eq. (6), $\mathbf{v_u}$ and $\mathbf{v}$ denote the labels predicted by the trainer and the trainee models respectively. Further, c_m represents the elements of the vector $\mathbf{c} \in \mathbb{R}^{257}$. It should be noted that M indicates the number of elements in $\mathbf{c}$ (for our case, $M = 257$).

3.2. Multi-Occlusion per Identity Network

In Fig. 1b, we present a novel end-to-end training pipeline. The details of our network are discussed below.

3.2.1 Network Architecture

Our training framework (see Fig. 1b) uses the ResNet-50 model as backbone architecture. The model takes 7 images per person (partially (or fully) occluded) with random patterns and returns 7 vectors of 3DMM coefficients with the dimension of 257. We exploit face recognition FaceNet [34], labels from Single-Occluded per Identity model from Sec. 3.1.1 to train the proposed model. Further, we modify the fully-connected classification layer of ResNet-50 with 7 fully-connected layers with 257 nodes for obtaining the coefficients denoted by $\mathbf{v_m} \in \mathbb{R}^{257 \times 7}$.

3.2.2 Losses

In this section, we detail the losses for training the model under discussion.

- **Perceptual Loss:** We introduce a perceptual loss to ensure the visual similarity *between* projected 3D face meshes (I'_m) obtained from multi-occlusion images of each identity. The loss serves as a weak-supervision signal from a pre-trained deep face recognition network [34]. Notably, we extract the deep features of the projected 3D face meshes and compute the perceptual loss between the projected faces of each identity:

$$L_{perc} = \frac{\sum_{q'=1}^{w_1} \sum_{q=1}^{w_2} \chi\chi^{\intercal}}{||\chi||_{2,1}||\chi^{\intercal}||_{2,1}}. \tag{7}$$

  In Eq. 7, χ is the deep feature matrix (containing feature vectors for 7 projected 3D faces) obtained from face recognition network. Also, q and q' represent rows and columns of a matrix respectively whereas $w_1 = w_2 = 7$.

- **Label Loss:** We exploit labels from the SOPI for the set of images I_m to the train the proposed model. We choose the SOPI label for the training such that the perceptual error of corresponding rendered image (obtained from the SOPI model) is minimum with the unoccluded face image. The loss enforces the model to restrict the search in 3DMM space and learn a single deep representation for the 3D face meshes obtained from multiple variations of the same face image.

$$\begin{aligned} L_{label} &= \sum_{j=1}^{7} ||{\mathbf{v_s}}^i - {\mathbf{v_m}}^j||_2, \quad \text{where} \\ {\mathbf{v_s}}^i &: \min \left(1 - \frac{\langle {\mathbf{f_s}}^i, \mathbf{f} \rangle}{||{\mathbf{f_s}}^i||||\mathbf{f}||} \right). \end{aligned} \tag{8}$$

  In Eq. 8, $\mathbf{f_s}$ and $\mathbf{f}$ are the deep-features derived from FaceNet model [34] corresponding to rendered images

obtained from SOPI and the unoccluded face image respectively. We choose i-th coefficient derived from SOPI ($\mathbf{v_s}^i$) such that the cosine distance ($\langle \cdot, \cdot \rangle$) between i-th feature $\mathbf{f_s}^i$ and $\mathbf{f}$ is minimum. Therefore, L_{label} (label loss) is derived between the 7 predicted labels $\mathbf{v}_m$ and the label $\mathbf{v_s}^i$.

Composite Multi-Occlusion Loss: To enforce the learning of the proposed model, we use Eqs. 7 and 8 to propose the novel composite loss function.

$$L = x_1 \times L_{perc} + x_2 \times L_{label}. \tag{9}$$

Here, we use $x_1 = 0.2$ and $x_2 = 0.8$ are the weights corresponding to the respective losses.

4. Datasets and Results

We gather *in-the-wild* face images from the *training and testing* datasets of CelebA released by Tensorflow [28]. Thus a major proportion of the dataset contains *naturally* occluded face images. Moreover, the dataset also includes face images of individuals under distinct illumination conditions and poses. It contains images of prominent persons from different races. Hence, we choose this *equitable* dataset with $\sim$181K images for training our model. We extract a quarter of the images from the dataset. We partially or fully (artificially) occlude these face images with arbitrary pixel values. We use $\sim$45K occluded face images for training our Single-Occlusion per Identity model. Besides, we use 7 random occlusion patterns to occluded each face image in the dataset to train the Multi-Occlusion per Identity model. The varying pixel values of the occluded region force the proposed models to learn without developing pixel-value bias.

Further, we provide the input images of size 224×224 to both models under discussion (see Fig. 1). We initialize our models with Imagenet weights [33]. Our models use Adam [23] optimizer with an initial learning rate as 10^{-5}, batch size of 5. The Single-Occlusion per Identity model requires 400K iterations, whereas the Multi-Occlusion per Identity model uses 500K iterations for the training.

4.1. Qualitative Results

Using the proposed approaches, we obtain the 3D faces for various comparisons. Our Single-Occlusion per Identity method exploits RNet from [10] to train the respective framework. Hence, we rigorously compare our 3D faces for occluded images with the outputs of RNet.

We also compare the performance of the Multi-Occlusion per Identity model with the Single-Occlusion per Identity approach, MoFA, and RNet.

4.1.1 Single-Occlusion per Identity

We use square-shaped occlusion patterns for the comparison as the area of a square is more than any other standard pattern. Thus square-shaped patterns enable larger area to be occluded in an image. We divide the occlusions into several levels to establish the efficacy of the proposed model. The face images with two occluded areas, covering a smaller region of the face in the image (see Fig. 2a, row 2), we refer to as Double Slightly Occluded Face Images (*DSOFI*). Besides, face images with double (see Fig. 2b, row 2) and single (see Fig. 2c, row 2) heavily occluded area(s) are Double Heavily Occluded Face Images (*DHOFI*) and Single Heavily Occluded Face Images (*SHOFI*) respectively. We perform the analysis on square-shaped (and/or rectangular) occlusion patterns for DSOFI, DHOFI and SHOFI to obtain a common ground for comparison of the network performance on various levels of occlusion in a face image.

Comparison for DSOFI: We compare the results for the DSOFI with the output of RNet. In Fig. 2a, our model produces a highly reliable output. Besides, RNet reconstructs 3D faces with random poses, shapes, expressions, and textures.

Comparison for DHOFI: Further, we compare our output for DHOFI with the 3D faces obtained using RNet. Fig. 2b shows that RNet reconstructs the *ghost-like* 3D face with generally no similarity with human faces. In contrast, our model produces output with a high similarity towards the corresponding unoccluded input image.

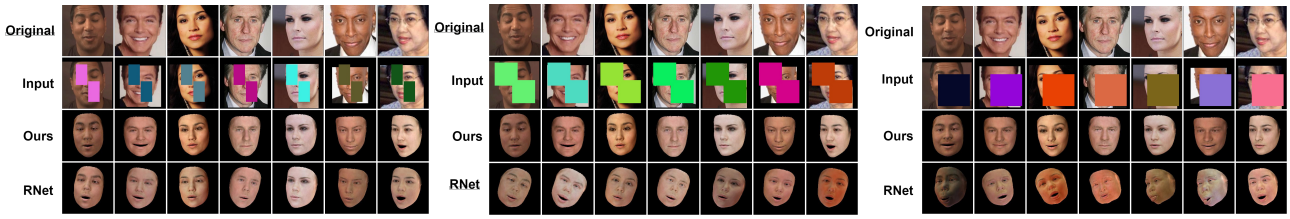
Comparison for SHOFI: We compare our output for SHOFI images with the results of RNet. We observe that (see Fig. 2c) RNet reconstructs the same *ghost-like* 3D face for all the heavily occluded input images. On the other hand, our model produces output with a likeness towards the corresponding 2D unoccluded face image. Note that we *estimate* the *probable* face shape and texture behind the occlusion. Thus *reducing the dimension* of the search space of 3D face recognition systems.

Impact of Occlusion Color on 3D Faces: We compare the impact of occlusions with different pixel-values on the 3D reconstructed face mesh with RNet. In Fig. 3, we observe that the skin tone of *ghost-like* faces in row 3 changes with the change in the pixel-value of occlusion. Contrarily, our model (see row 2 of Fig. 3) is resistant to the color of occlusion. We can infer that our model possesses differentiation ability between skin tone and occlusion color.

Impact of Real-life Occlusions on 3D Faces: To illustrate the efficacy of the proposed method for tackling the real-life occlusions, we perform a qualitative comparison of our method with RNet and MofA [40]. Fig. 4 shows that the proposed model performs better than the state-of-the-art methods.

4.1.2 Multiple-Occlusion per Identity

We also perform qualitative analysis on the multi-occlusion per identity model. We show the impact of various patterns of occlusions on the 3D reconstructed face mesh obtained



(a) DSOFI. (b) DHOFI. (c) SHOFI.

Figure 2: A quantitative analysis on the 7 images from the MoFA dataset [40].

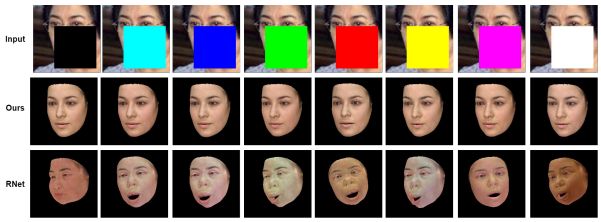


Figure 3: A qualitative illustration of the ability of our model to differentiate between the skin tone and the occlusion color. We obtained the images from MoFA-test dataset.

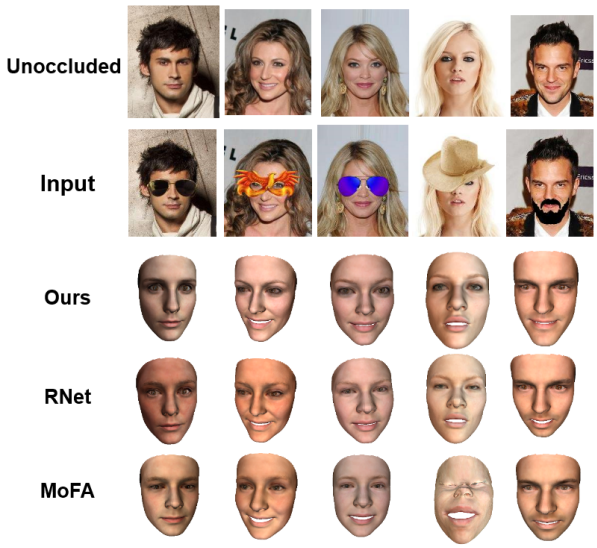


Figure 4: Our model to reconstructs reliable 3D faces from real-life occlusions on the face images.

using the RNet, Multi-Occlusion per Identity, and Single-Occlusion per Identity models. Fig. 5 emphasizes that our Multi-Occlusion per Identity model possesses the highest 3D face reconstruction accuracy for arbitrary (patterns) occluded face images. The proposed model is restricted to reconstruct 3D faces from artificially occluded face images only (due to the dataset limitation). Thus we compare our results for the same category of occlusion.

4.2. Quantitative Analysis

We quantitatively evaluate our results on two publicly available datasets: validation dataset of CelebA released by Tensorflow [28] and 34 images of AFLW2000-3D dataset [46]. Thus, we validate our models using $\sim$ 19K images for face occlusions. Note that we *do not* use validation dataset of CelebA for training the model. Hence the dataset is suitable for performing the quantitative analysis of our models. We compare the landmark accuracy between our results and RNet in Sec. 4.2.1 and 4.2.2. Further, we also compare the texture accuracy of our results with the outputs of the RNet model in Secs. 4.2.1 and 4.2.2. We use histogram dissimilarity metric for the comparison.

4.2.1 Single-Occlusion per Identity

In this section, we compare the results of the Single-Occlusion per Identity model with RNet. Note that we perform a comparison for landmark accuracy and histogram dissimilarity of the results.

Comparison on the Landmark Accuracy: We compare the landmark accuracy of our model with RNet. We use dlib landmark detector [22] to evaluate 68 facial landmarks. In Fig. 2a, we observe that RNet fails to reconstruct reliable face poses. Thus, landmark accuracy is in jeopardy. Moreover, in Fig. 2b and 2c, RNet reconstructs *ghost-like* faces. Thus, we compare the landmark accuracy of RNet for unoccluded images only. We evaluate the landmark accuracy of our model under several occlusions in Tables 1 and 3. We compare the averages of the difference in the euclidean distance (ME) and the squared mean of the difference between landmark coordinates of the input image and the reconstructed 3D face (MSE) for the validation dataset of CelebA. Further, we evaluate *normalized* (landmark coordinate divided by 224) euclidean (normalized ME) and MSE landmark errors (normalized MSE) for AFLW2000-3D dataset (see Table 1). Note that the landmark accuracy of our model is comparable to the RNet for the unoccluded images. Moreover, the landmark accuracy for DSOFI, DHOFI, and SHOFI is as precise as the landmark accuracy for the unoccluded faces. In the Table 1, $--$ implies *no reliable results obtained*.

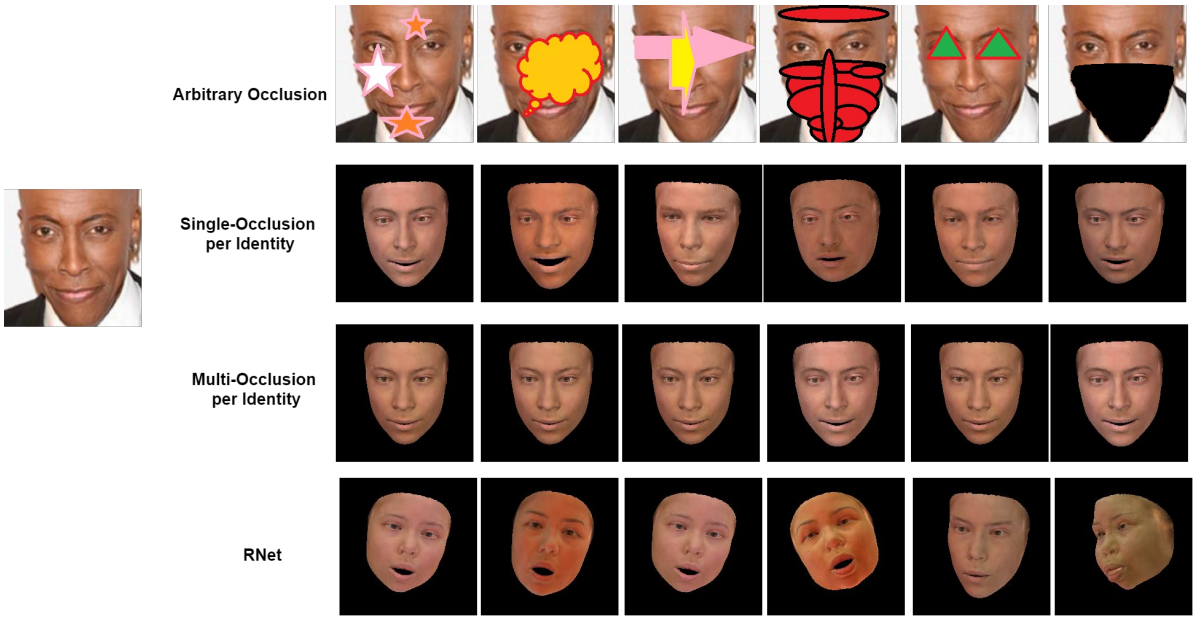


Figure 5: Our Multi-Occlusion per Identity model reconstructs reliable 3D faces from arbitrary patterns of occlusions. The top row contains occluded face images.

Occlusion	CelebA				AFLW2000-3D			
	ME (↓)		MSE (↓)		Normalized ME (↓)		Normalized MSE (↓)	
	Ours	RNet	Ours	RNet	Ours	RNet	Ours	RNet
Unoccluded	1.24	1.24	(3.25, 1.66)	(3.23, 1.69)	0.68	0.68	(0.26, 0.43)	(0.26, 0.43)
DSOFI	1.25	−−	(3.27, 1.68)	−−	0.68	−−	(0.26, 0.43)	−−
DHOFI	1.24	−−	(3.20, 1.66)	−−	0.68	−−	(0.26, 0.43)	−−
SHOFI	1.26	−−	(3.26, 1.70)	−−	0.69	−−	(0.26, 0.44)	−−

Table 1: A comparison of the landmark errors of our model with RNet on validation dataset of CelebA released by Tensorflow and AFLW2000-3D dataset. We compare the landmarks on two criteria: **1)** average of differences in euclidean distances between facial landmarks of input images and the obtained meshes, and **2)** average mean square error separately for x and y coordinates of landmark points between input face images and the corresponding meshes. We performed the comparison for unoccluded images, DSOFI, DHOFI, and SHOFI.

Occlusion	CelebA		AFLW2000-3D	
	Ours	RNet	Ours	RNet
Unoccluded	4.76	4.74	3.82	3.81
DSOFI	4.77	4.90	3.82	4.02
DHOFI	4.79	7.98	3.84	5.36
SHOFI	4.83	−−	3.88	−−

Table 2: A comparison of histogram dissimilarity (↓) of our model with RNet on validation dataset of CelebA released by Tensorflow and AFLW2000-3D dataset.

Comparison of Histogram Dissimilarity: In this section, we compare the histogram dissimilarity (texture-based analysis) of our results with the outputs of the RNet model. In Tables 2 and 4, we observe that our model performs better than the RNet model for various proposed categories of occlusions viz, DSOFI, DHOFI and SHOFI, and natural cases. Note that the RNet performs poorly for SHOFI. Hence we do not provide the histogram similarity for the SHOFI.

Method	Landmark Inaccuracy (ME ↓)
RNet	5.84
Ours	**0.77**

Table 3: A comparison of landmark inaccuracy (ME) of our model with RNet on 54 *naturally* occluded images from the validation dataset of CelebA released by Tensorflow.

Method	Histogram Dissimilarity (↓)
RNet	5.01
Ours	**4.76**

Table 4: A comparison of histogram dissimilarity (↓) of our model with RNet on 54 *naturally* occluded images from the validation dataset of CelebA released by Tensorflow.

4.2.2 Multi-Occlusion per Identity

We compare the results of the Multi-Occlusion per Identity model with the RNet. Following Sec. 4.2.1, we perform comparisons on the landmark accuracy and histogram dissimilarity metric. Note that unlike the evaluation of single-occlusion per identity, we do not fix the occlusion patterns viz, SHOFI, DHOFI, and DSOFI. Instead, we use random patterns of occlusions across the evaluation dataset.

Comparison on the Landmark Accuracy: We compare the landmark accuracy for random patterns on the images. Our results in Table 5 demonstrates superior performance than the RNet.

Method	CelebA		AFLW2000-3D	
	ME (↓)	MSE (↓)	Normalized ME (↓)	Normalized MSE (↓)
RNet	7.91	(5.65, 3.94)	0.87	(0.64, 0.93)
Ours	1.24	(3.21, 1.67)	0.68	(0.26, 0.44)

Table 5: A comparison of landmark errors of our Multi Occlusion per Identity model with the RNet on validation dataset of CelebA released by Tensorflow and AFLW2000-3D dataset. We compare the accuracy on random patterns such as stars, crown (no specific standard shape).

Method	CelebA	AFLW2000-3D
RNet	6.93	4.74
Ours	4.76	3.82

Table 6: A comparison of histogram dissimilarity (↓) of our model with RNet on validation dataset of CelebA and AFLW2000-3D dataset.

Comparison of Histogram Dissimilarity: In this section, we compare our results with the RNet using the histogram similarity metric. In Table 6, we demonstrate the superior performance of our model than the RNet for random patterns of occlusions.

4.3. Ablation Study

In this section, we validate the efficacy of the SE loss function for training our Single-Occlusion per Identity model. Further, we demonstrate the effectiveness of various components of the proposed composite multi-occlusion loss in training the Multi-Occlusion per Identity model.

4.3.1 Ablations on Single-Occlusion per Identity

In this section, we perform ablations on the various potential losses for training Single-Occlusion per Identity model.

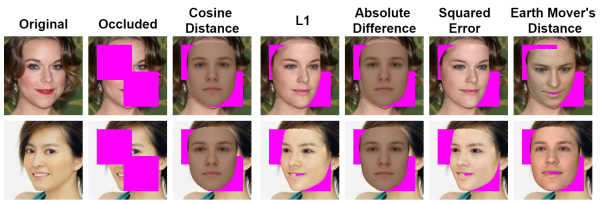


Figure 6: Impact of various losses on the training of proposed SOPI model. The *original* images are obtained from the CelebA dataset.

We qualitatively examine the outputs of the models thus obtained. Fig. 6 demonstrates the preliminary results emphasizing on the efficacy of SE and L1 loss function for training the proposed SOPI model.

4.3.2 Ablation study Multi-Occlusion per Identity

We perform an ablation study on various losses of the proposed novel composite multi-occlusion loss. We compare the performance of the model based on landmark accuracy. Table 7 demonstrates the effectiveness of the components of various losses in the proposed loss function.

L_{perc}	L_{label}	CelebA		AFLW2000-3D	
		ME (↓)	MSE (↓)	Normalized ME (↓)	Normalized MSE (↓)
✓		41.0	(32.84, 23.97)	0.94	(0.75, 0.88)
	✓	2.11	(3.10, 2.23)	0.71	(0.32, 0.46)
✓	✓	1.24	(3.21, 1.67)	0.68	(0.26, 0.44)

Table 7: An ablation study on Multi-Occlusion per Identity model. We compare the landmark accuracy on random patterns such as stars, crown (no specific shape).

5. Conclusion

We have shown that our **S**ingle-**O**cclusion **P**er **I**dentity (**SOPI**) model reconstructs accurate 3D faces from occluded face images. We demonstrated the efficacy of the proposed method using three variations of artificially occluded face data: SHOFI, DSOFI, and DHOFI. Further, we showed that the proposed model is invariant towards the color of the occlusions in the face images. Our SOPI model obtained a landmark accuracy of 0.77 against 5.84 of recent state-of-the-art-method for real-life challenging facial occlusions. Also, we presented a novel **M**ulti-**O**cclusion **P**er **I**dentity (**MOPI**) model to emphasize the significance of visible facial features during learning. Our model demonstrated a reduction in the landmark error by a margin of 6.67 compared to the state-of-the-art method for artificial occlusions. We have shown superior performance of the MOPI model towards face occlusions than the SOPI network. Besides, we have illustrated the lesser data requirement and training simplicity of SOPI model over MOPI approach.

References

[1] 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.

[2] Anil Bas, William AP Smith, Timo Bolkart, and Stefanie Wuhrer. Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision*, pages 377–391. Springer, 2016.

[3] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *Computer graphics forum*, volume 22, pages 641–650. Wiley Online Library, 2003.

[4] Volker Blanz, Albert Mehl, Thomas Vetter, and H-P Seidel. A statistical method for robust 3d surface reconstruction from sparse data. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 293–300. IEEE, 2004.

[5] 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.

[6] 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.

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[8] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9429–9439, 2019.

[9] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[10] 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.

[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] 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.

[13] Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. Perspective-aware manipulation of portrait photos. *ACM Transactions on Graphics (TOG)*, 35(4):1–10, 2016.

[14] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):1–15, 2016.

[15] 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.

[16] 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.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[18] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[19] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1294–1307, 2018.

[20] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4295–4304, 2015.

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[22] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[23] Diederik P Kingma and Jimmy Ba. Adam: A methodfor stochastic optimization. In *International Conference on-Learning Representations (ICLR)*, 2015.

[24] Martin D Levine and Yingfeng Chris Yu. State-of-the-art of 3d facial reconstruction methods for face recognition based on a single 2d training image per person. *Pattern Recognition Letters*, 30(10):908–913, 2009.

[25] 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.

[26] Jiangke Lin, Yi Yuan, and Zhengxia Zou. Meingame: Create a game character face from a single portrait. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[27] 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.

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738. IEEE Computer Society, 2015.

[29] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[30] 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.

[31] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 986–993. IEEE, 2005.

[32] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, 2016.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[34] 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.

[35] 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.

[36] 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.

[37] Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M Seitz. Total moving face reconstruction. In *European conference on computer vision*, pages 796–812. Springer, 2014.

[38] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images, 2020.

[39] 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.

[40] 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.

[41] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, pages 3935–3944, 2018.

[42] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018.

[43] 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.

[44] Jialiang Zhang, Lixiang Lin, Jianke Zhu, and Steven C. H. Hoi. Weakly-supervised multi-face 3d reconstruction, 2021.

[45] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[46] 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.

[47] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 787–796, 2015.