electronics

MDPI

Article

Convincing 3D Face Reconstruction from a Single Color Image under Occluded Scenes

Dapeng Zhao [1], Jinkang Cai [2] and Yue Qi [1,3,4,*]

[1] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100190, China; mirror1775@gmail.com
[2] School of Transportation Science and Engineering, Beihang University, Beijing 100190, China; jinkangcai@buaa.edu.cn
[3] Peng Cheng Laboratory, Shenzhen 518066, China
[4] Qingdao Research Institute, Beihang University, Qingdao 266000, China
* Correspondence: qy@buaa.edu.cn

Abstract: The last few years have witnessed the great success of generative adversarial networks (GANs) in synthesizing high-quality photorealistic face images. Many recent 3D facial texture reconstruction works often pursue higher resolutions and ignore occlusion. We study the problem of detailed 3D facial reconstruction under occluded scenes. This is a challenging problem; currently, the collection of such a large scale high resolution 3D face dataset is still very costly. In this work, we propose a deep learning based approach for detailed 3D face reconstruction that does not require large-scale 3D datasets. Motivated by generative face image inpainting and weakly-supervised 3D deep reconstruction, we propose a complete 3D face model generation method guided by the contour. In our work, the 3D reconstruction framework based on weak supervision can generate convincing 3D models. We further test our method on the MICC, Florence and LFW datasets, showing its strong generalization capacity and superior performance.

Keywords: 3D face reconstruction; face parsing; occluded scenes

Citation: Zhao, D.; Cai, J.; Qi, Y. Convincing 3D Face Reconstruction from a Single Color Image under Occluded Scenes. *Electronics* **2022**, *11*, 543. https://doi.org/10.3390/electronics11040543

Academic Editor: Jeha Ryu

Received: 11 January 2022
Accepted: 4 February 2022
Published: 11 February 2022

1. Introduction

Single-view 3D face reconstruction refers to obtaining a user-specific 3D face surface model given one input face image. This is a classical and fundamental problem in computer vision [1–3]. It has a wide range of applications, such as 3D-assisted face recognition [1,4–6] and digital entertainment [7]. Existing methods mainly concentrate on reconstructing beautiful textures and ignore geometric details. At the same time, these methods can only work effectively when frontal faces are unobstructed, which makes the application of scenes very limited. When considering the occlusion of the scene, the reconstruction of the 3D face model is challenging since part of the facial features is not visible.

In recent years, due to the rapid development of deep learning methods, similar face inpainting tasks have made significant breakthroughs [8,9]. By comparison, because deep learning methods cannot be applied to 3D structures end-to-end, 3D reconstruction methods have remained far behind [10].

In 1999, Blanz and Vetter proposed early 3D morphable models (3DMM) [1,11,12], and the field of 3D face reconstruction using a single image opened. These approaches are based on automated template matching robust reproduction results within 1–5 min. However, due to the constraint space's existence, the model's performance is still lacking in competitiveness in terms of the expressiveness of geometric details [13]. At the same time, 3DMM and other methods also cannot deal with the situation in which faces occlude scenes robustly (especially the texture). These methods generally indiscriminately reconstruct the occluded field of face. Unlike previous arts, we propose a method designed to attain both goals: detailed 3D face reconstruction and robustness to occlusions (Figure 1). How did

we do it? With the assistance of the face parsing framework, face contour map and deep learning method, we find a way to identify the occluded area and reconstruct the input image to an accurate 3D face model.

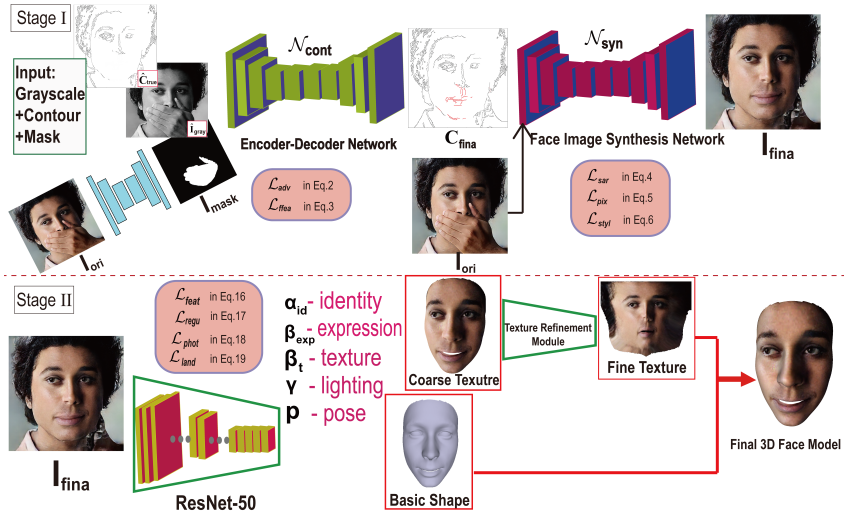

Figure 1. Method overview. See related sections for details.

The main contributions are summarized as follows:

- We propose a novel approach that combines the face parsing approach and face contour map to generate a face with complete facial features.
- Face occlusion is a common problem. In response to the problem of an invisible face area under occluded scenes, we propose synthesizing the input face image based on GANS rather than reconstructing the 3D face directly.
- We improved the loss function of our 3D face reconstruction framework for occluded scenes. Our results (especially the face texture) are more accurate than other recent methods.

2. Related Work

2.1. Single-View 3D Face Shape Prediction

When it comes to 3D face reconstruction, the classic methods use reference 3D face models to fit the input face photo. The first step is face alignment. Face alignment, which fits a face model to an image and extracts the fiducial facial landmarks, has many solutions in the CV community. These solutions including the active appearance model [14–16] and the constrained local model [17–19]. Besides traditional models, some recent techniques use convolutional neural networks (CNNs) to regress landmark locations with the raw face image [20–22].

The second step is to solve the nonlinear optimization function to regress the 3DMM coefficients [1]. Some recent techniques firstly used CNNs to predict the 3DMM parameters with the input face image [2,23,24]. Some works proposed a cascaded CNN structure to regress the accurate 3DMM shape parameters [25–29]. Some frameworks explored the end-to-end CNN architectures to regress 3DMM coefficients directly. Each calculation usually takes a long time because the dimensionality of the data is very high [30].

2.2. Face Parsing

A face parsing map generally serves as an intermediate representation for conditional face image generation [31]. In addition, the image-to-image GAN model can learn the mapping from the semantic map to realistic RGB image [32–35]. In the pixel-level image semantic segmentation methods based on deep learning, fully convolutional networks (FCN) [36] is the well-known baseline for generic images which analyze per-pixel feature. Following this work, the DeepLab approaches [37–40] have achieved impressive results.

The main feature of the series is to use dilated convolution instead of traditional convolution. However, directly applying these frameworks for face parsing may fail to map the varying-yet-concentrated facial features, especially hair, leading to poor results. A workable solution should directly predict per-pixel semantic label across the entire face photo. Wei et al. [41] proposed a novel method for regulating receptive fields with superior regulation ability in parsing networks to access accurate parsing map. MaskGan [42] contributed a labeled face dataset [43]. Zhou et al. [44] proposed an architecture that explores how to combine the fully convolutional network model and super-pixel data to model together. In order to solve the question of global image information access restriction, some methods [45] have introduced the transformer component and achieved state-of-the-art results. The semantic layout guides the location and appearance of facial features and further facilitates the training. The majority of face parsing methods work require semantic labels. Hence, these frameworks [42,46–51] usually train on the CelebA and Helen datasets, which contain labeled attributes.

2.3. Generative Adversarial Networks

Generative adversarial networks (GANs) [52] generally consist of a generator and a discriminator. The two components compete with each other. Since GANs can generate realistic images, GANs have been successfully applied to various face image synthesis tasks, such as image manipulation [53], image-to-image translation [54], image inpainting [55] and texture blend [56–58]. For example, the face images generated by Stylegan2 [59] can be confused with the real. With continuous improvements in regularization [60], users can control the synthesis by feeding the generator with conditioning information instead of noise. Our work was built on conditional GANs [61] with face parsing map inputs, which aims to tackle facial reconstruction under occluded scenes.

2.4. Face Image Synthesis

Deep pixel-level face generating has been studied for several years. Many methods [46,62–65] have achieved remarkable results. Context encoder [66] is the first deep learning network designed for image inpainting with the encoder–decoder architecture. Nevertheless, the networks do a poor job in dealing with human faces. Following this work, Yang et al. [35] used a modified VGG network [67] to improve the result of the context encoder by minimizing the feature difference of the photo background. Dolhansky et al. [68] demonstrated the significance of exemplar data for inpainting. However, this method only focuses on filling in missing eye regions of the frontal face, so it does not generalize well. EdgeConnect [69] shows impressive proceeds, disentangling generation into two stages: edge generator and image completion network. Contextual attention [70] takes a similar two-step approach. First, it produces a base estimate of the invisible region. Next, the refinement block sharpens the photo by background patch sets. The typical limitations of current face image generate schemes are the necessity of manipulation, the complexity of fundamental architectures, the degradation in accuracy, and the inability of restricting modification to local region.

3. Our Method

We propose a detailed 3D face reconstruction method (as shown in Figure 1) based on a single photo that consists of two stages:

- In response to the occlusion area, synthesizing the 2D face with complete facial features.
- Detailed 3D shape reconstruction module based on unobstructed frontal images.

Our goal is to realize detailed 3D face shape reconstruction under occluded scenes using our method. Given a source color face image $\mathbf{I_{ori}} \in \mathbb{R}^{H \times W \times 3}$ with obstructions, we obtain the final 3D face model.

3.1. Face Mask Generation

As the first step of our 3D reconstruction framework, we need to identify the occluded area for generating the face mask image (1 for the occluded region, 0 for background) for the next task. Inspired by traditional face parsing tasks, as shown in Figure 2, given a square image $\mathbf{I_{ori}} \in \mathbb{R}^{H\times W\times 3}$ of the face under occlusion, we applied the trained face mask generator $\mathcal{N}_{mask}$ to obtain the face mask $\mathbf{I_{mask}} \in \mathbb{R}^{H\times W\times 1}$. This mask generation task is very similar to the traditional face parsing task. Our face mask generator is partly inspired by the annotated face dataset CelebAMask-HQ [42]. We trained an encoder–decoder module $\mathcal{N}_{mask}$ based on U-Net [71] to predict the occluded region.

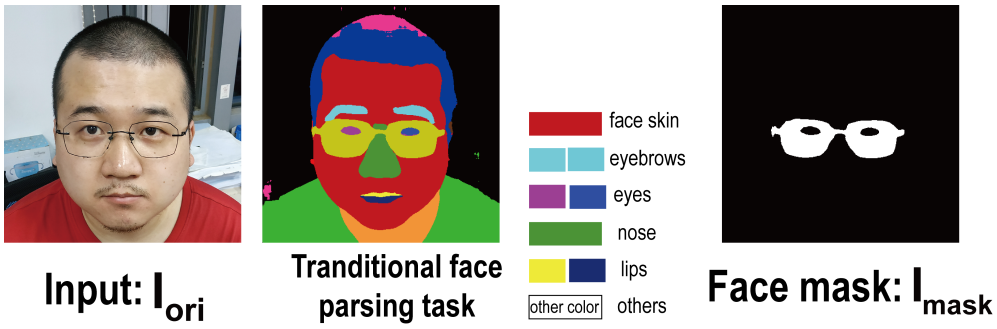

Figure 2. Our face mask generation module. It is slightly different from the traditional face parsing task. The traditional face parsing task is to recognize the face as different components (usually including eyebrows, eyes, nose, mouth, facial skin and so on). Corresponding to it is the face parsing map (different face components are represented by different gray values). Our mask generation task is only to recognize the occluded area. The corresponding face mask map is a binary map.

3.2. Face Image Synthesis with GANs

Our face image synthesis module is guided by the contour. First, we need to predict the contours $\mathbf{C_{syn}} \in \mathbb{R}^{H\times W\times 1}$ of facial features in the occluded area. We assume that the final unobstructed face image is $\mathbf{I_{fina}} \in \mathbb{R}^{H\times W\times 3}$ and the ground truth image without obstruction $\mathbf{I_{true}} \in \mathbb{R}^{H\times W\times 3}$. In the training set, the corresponding complete contour image and gray image are $\mathbf{C_{true}} \in \mathbb{R}^{H\times W\times 1}$ and $I_{\mathbf{gray}} \in \mathbb{R}^{H\times W\times 1}$. We trained the contour generator $\mathcal{N}_{cont}$ to predict the contour map for the occluded region.

$$\begin{cases} \mathbf{C_{syn}} = \mathcal{N}_{cont}(\hat{\mathbf{I}}_{\mathbf{gray}}, \mathbf{I_{mask}}, \hat{\mathbf{C}}_{\mathbf{true}}) \\ \hat{\mathbf{I}}_{\mathbf{gray}} = \mathbf{I_{gray}} \odot (1 - \mathbf{I_{mask}}) \\ \hat{\mathbf{C}}_{\mathbf{ture}} = \mathbf{C_{ture}} \odot (1 - \mathbf{I_{mask}}) \end{cases} \tag{1}$$

where $\hat{\mathbf{I}}_{\mathbf{gray}}$ denotes the masked grayscale image, $\hat{\mathbf{C}}_{\mathbf{ture}}$ denotes the masked contour image and $\odot$ denotes the Hadamard product. We trained the discriminator of the module $\mathcal{N}_{cont}$ to predict which of $\mathbf{C_{syn}}$ and $\mathbf{C_{ture}}$ is a true contour map and which is a false contour map. The adversarial loss is defined as

$$\mathcal{L}_{adv} = \mathbb{E}_{(\mathbf{C_{ture}}, \mathbf{I_{gray}})}\left[\log D_1(\mathbf{C_{true}}, \mathbf{I_{gray}})\right] + \mathbb{E}_{(\mathbf{I_{gray}})}\log\left[1 - D_1(\mathbf{C_{syn}}, \mathbf{I_{gray}})\right] \tag{2}$$

where $\mathbb{E}$ denotes the expected value of the function, and D_1 denotes the discriminator of the adversarial loss function.

In addition, we compare the feature activation maps of the discriminator. We set the face feature matching loss as

$$\mathcal{L}_{ffea} = \mathbb{E}\left[\sum_{i=1}^{K}\frac{1}{N_i}\left\|D_1^{(i)}(\mathbf{C_{true}}) - D_1^{(i)}(\mathbf{C_{syn}})\right\|\right] \tag{3}$$

where N_i is the number of elements in the ith activation layer, K is the final convolution layer of the discriminator and $D_1^{(i)}$ is the activation in the i-th layer of the discriminator.

After obtaining the complete contour map, we design $\mathcal{N}_{syn}$ to generate the complete face image $\mathbf{I_{fina}}$. The complete contour map $\mathbf{C_{fina}}$ is formed by adding $\mathbf{C_{syn}}$ and $\mathbf{C_{true}}$, which follows $\mathbf{C_{fina}}$=$\mathbf{C_{true}} \odot (1 - \mathbf{I_{mask}}) + \mathbf{C_{syn}} \odot \mathbf{I_{mask}}$. In the map $\mathbf{C_{fina}}$, we can see the contours of all facial features, especially the occluded areas. In addition, we set $\hat{\mathbf{I}}_{\mathbf{true}} \in \mathbb{R}^{H \times W \times 3}$ to be an incomplete face picture, which follows $\hat{\mathbf{I}}_{\mathbf{true}} = \mathbf{I_{true}} \odot (1 - \mathbf{I_{mask}})$. So, we utilize $\mathcal{N}_{syn}$ to obtain the final complete face image $\mathbf{I_{fina}}$, with occluded regions recovered, which follows $\mathbf{I_{fina}}$=$\mathcal{N}_{syn}(\hat{\mathbf{I}}_{\mathbf{true}}, \mathbf{C_{fina}})$.

We trained the module $\mathcal{N}_{syn}$ to predict the final complete face image $\mathbf{I_{fina}}$ over a joint loss. The adversarial loss is defined as

$$\mathcal{L}_{sar} = \mathbb{E}_{(\mathbf{I_{true}}, \mathbf{C_{fina}})}[\log D_2(\mathbf{I_{true}}, \mathbf{C_{fina}})] + \mathbb{E}_{(\mathbf{C_{fina}})} \log[1 - D_2(\mathbf{I_{fina}}, \mathbf{C_{fina}})] \tag{4}$$

The per-pixel loss [72] is defined as follows:

$$\mathcal{L}_{pix} = \frac{1}{S_m} \|\mathbf{I_{fina}} - \mathbf{I_{true}}\|_1 \tag{5}$$

where S_m denotes the size of the face mask $\mathbf{I_{mask}}$, and $\|\cdot\|_1$ denotes the L_1 norm. Notice that we use the mask size S_m as the denominator to adjust the penalty.

The style loss [73] computes the style distance between two face images as follows

$$\mathcal{L}_{styl} = \sum_n \frac{1}{Q_n \times Q_n} \left\| \frac{G_n(\mathbf{I_{fina}} \odot (1 - \mathbf{I_{mask}})) - G_n(\hat{\mathbf{I}}_{\mathbf{true}})}{Q_n \times H_n \times W_n} \right\|_1 \tag{6}$$

where $G_n(x) = \varphi_n(x)^T \varphi_n(x)$ denotes the gram matrix corresponding to $\varphi_n(x)$, and $\varphi_n(\cdot)$ denotes the Q_n feature maps with the size $H_n \times W_n$ of the n-th layer.

In summary, the contour generator network $\mathcal{N}_{cont}$ was trained with an objective comprised of an adversarial loss and feature-matching loss

$$\min_{G_1} \max_{D_1} \mathcal{L}_{G_1} = \min_{G_1} \left(\lambda_{adv} \max_{D_1} \mathcal{L}_{adv} + \lambda_{ffea} \mathcal{L}_{ffea} \right) \tag{7}$$

The total loss function of $\mathcal{N}_{syn}$ follows

$$\min_{G_2} \max_{D_2} \mathcal{L}_{G_2} = \lambda_{sar} \max_{D_2} \mathcal{L}_{sar} + \lambda_{pix} \mathcal{L}_{pix} + \lambda_{styl} \mathcal{L}_{styl} \tag{8}$$

where we set $\lambda_{adv} = 1$, $\lambda_{ffea} = 11.5$, $\lambda_{sar} = 0.1$, $\lambda_{pix} = 1$ and $\lambda_{styl} = 250$, respectively. The values of these weights refer to the method of Lee et al. [42].

3.3. 3D Shape Model

A 3DMM consists of three model parts: the shape, texture and camera models. Let us denote the 3D shape and texture of an object with n vertices as a $3n \times 1$ vector:

$$\mathbf{S} = (x_1, y_1, z_1, \ldots, x_n, y_n, z_n) \tag{9}$$

$$\mathbf{T} = (r_1, g_1, b_1, \ldots, r_n, g_n, b_n) \tag{10}$$

where $\mathbf{S_i} = (X_i, Y_i, Z_i)$ denotes the object-centered shape vector of the i-th vertex, and $\mathbf{T_i} = (r_i, g_i, b_i)$ denotes the texture vector of the i-th vertex.

The face model to be solved can be weighted and combined by the m face model in the dataset:

$$\begin{cases} \mathbf{S_{mod}} = \sum_{i=1}^{m} \boldsymbol{\alpha_i} \mathbf{S_i} \\ \mathbf{T_{mod}} = \sum_{i=1}^{m} \boldsymbol{\beta_i} \mathbf{T_i} \\ \sum_{i=1}^{m} \boldsymbol{\alpha_i} = \sum_{i=1}^{m} \boldsymbol{\beta_i} = 1 \end{cases} \tag{11}$$

where $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ denote the weighting coefficient of the face model.

However, the basis vectors here are not orthogonally related. We normally use the following formula when building the model:

$$\mathbf{S_{mod}} = \overline{\mathbf{S}} + \sum_{i=1}^{m-1} \widetilde{\boldsymbol{\alpha}_\mathbf{i}} \widetilde{\mathbf{S}_\mathbf{i}}, \mathbf{T_{mod}} = \overline{\mathbf{T}} + \sum_{i=1}^{m-1} \widetilde{\boldsymbol{\beta}_\mathbf{i}} \widetilde{\mathbf{T}_\mathbf{i}} \tag{12}$$

where $\overline{\mathbf{S}}$ and $\overline{\mathbf{T}}$ denote the average shape and average texture, $\widetilde{\boldsymbol{\alpha}_\mathbf{i}} \in \mathbb{R}^{80}$, $\widetilde{\boldsymbol{\beta}_\mathbf{i}} \in \mathbb{R}^{80}$ denote the eigenvalue of the covariance matrix arranged in descending order by the value, and $\widetilde{\mathbf{S}}, \widetilde{\mathbf{T}}$ denote the eigenvector of the shape and texture covariance matrix.

In fact, only the first few components of $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{T}}$ need to be selected to make a better approximation to the face sample. Not only can the number of parameters that needs to be estimated be greatly reduced, but the accuracy will not be significantly reduced. We describe the basic 3D face space with PCA:

$$\mathbf{S_{basi}} = \overline{\mathbf{S}} + \mathbf{A_{id}} \boldsymbol{\alpha_{id}} + \mathbf{B_{exp}} \boldsymbol{\beta_{exp}}, \mathbf{T} = \overline{\mathbf{T}} + \mathbf{B_t} \boldsymbol{\beta_t} \tag{13}$$

where $\mathbf{A_{id}}$, $\mathbf{B_{exp}}$ and $\mathbf{B_t}$ denote the PCA bases of identity, expression and texture, $\boldsymbol{\alpha_{id}} \in \mathbb{R}^{80}$ and $\boldsymbol{\beta_{exp}} \in \mathbb{R}^{64}$, and $\boldsymbol{\beta_t} \in \mathbb{R}^{80}$ are the corresponding 3DMM coefficient vectors. We adopt the Basel Face Model (BFM) [12]. It is a publicly available 3DMM dataset for a single view face model.

3.4. Camera and illumination model

After the 3D face is reconstructed, it can be projected onto the image plane with the perspective projection

$$V_{2d}(\mathbf{P}) = f * \mathbf{P_r} * \mathbf{R} * \mathbf{S_{mod}} + \mathbf{t_{2d}} \tag{14}$$

where $V_{2d}(\mathbf{P})$ denotes the projection function that turned the 3D model into 2D face positions, f denotes the scale factor, $\mathbf{P_r}$ denotes the projection matrix, $\mathbf{R} \in SO(3)$ denotes the rotation matrix, and $\mathbf{t_{2d}} \in \mathbb{R}^3$ denotes the translation vector.

Therefore, we approximated the scene illumination with spherical harmonics (SH) [74–77] parameterized by coefficient vector $\gamma \in \mathbb{R}^9$. In summary, the unknown parameters to be learned can be denoted by a vector $y = (\boldsymbol{\alpha_{id}}, \boldsymbol{\beta_{exp}}, \boldsymbol{\beta_t}, \gamma, \mathbf{p}) \in \mathbb{R}^{239}$, where $\mathbf{p} \in \mathbb{R}^6 = \{\mathbf{pitch}, \mathbf{yaw}, \mathbf{roll}, f, \mathbf{t_{2D}}\}$ denote face poses. In this work, we used a fixed ResNet-50 [78] network to regress these coefficients. We used a coarse-to-fine network based on the graph convolutional networks of Lin et al. [79] for producing the fine texture $\mathbf{T_{fin}}$.

3.5. Loss Function of Shape Reconstruction

Given a synthetic face photo $\mathbf{I_{fina}}$, we used the ResNet to regress the corresponding coefficient y. Because the collection of large scale high-resolution 3D texture datasets is still very costly and scarce, the ResNet was trained under weak supervision. The corresponding loss function consists of four parts:

$$\mathcal{L}_{shape} = \lambda_{feat} \mathcal{L}_{feat} + \lambda_{regu} \mathcal{L}_{regu} + \lambda_{phot} \mathcal{L}_{phot} + \lambda_{land} \mathcal{L}_{land} \tag{15}$$

The second term is a regularizer, and the other terms are data terms. We used fixed $\lambda_\bullet$ values to weigh the losses. Here, we set λ_{feat}= 0.2,$\lambda_{regu} = 3.6e-4$,$\lambda_{phot} = 1.4$, $\lambda_{land} = 1.6e-3$, respectively, in all our experiments. The values of these weights refer to the method of Deng et al. [77].

Face Features Level Consistency [77,79,80]. Face recognition is a very mature research area. In order to measure the difference between the 3D face and the two-dimensional face, we introduced the loss at face features level. The face features level consistency measures the difference between the 2D input image $\mathbf{I_{fina}}$ and rendered image $\mathbf{I_j}$.

$$\mathcal{L}_{feat} = 1 - \frac{< F(\mathbf{I_{fina}}), F(\mathbf{I_j}) >}{\|F(\mathbf{I_{fina}})\| \cdot \|F(\mathbf{I_j})\|} \tag{16}$$

where $F(\cdot)$ denotes the feature extraction function by FaceNet [81], and $< \cdot, \cdot >$ denotes the inner product.

Regularization Consistency [77]. To prevent shape deformation, we introduce the prior distribution to the parameters of the 3DMM face model. We add the regularization consistency on the regressed 3DMM coefficients.

$$\mathcal{L}_{regu} = \omega_\alpha \|\widetilde{\boldsymbol{\alpha}}_\mathbf{i}\|^2 + \omega_\beta \left\|\widetilde{\boldsymbol{\beta}}_\mathbf{i}\right\|^2 \tag{17}$$

here, we set ω_α=1.0 ,$\omega_\beta = 1.75e-3$ respectively.

Photometric Consistency [11,82–84]. As a common weak supervision method, it is easy to think of the dense photometric discrepancy. The rendering module renders back an image $\mathbf{I}_\mathbf{j}^{(i)}$ to compare with the image $\mathbf{I}_{\mathbf{fina}}^{(i)}$.

$$\mathcal{L}_{phot}(y) = \frac{\sum_{i \in \Psi} Z_i \cdot \left\| I_{\mathbf{fina}}^{(i)} - I_\mathbf{j}^{(i)} \right\|_2}{\sum_{i \in \Psi} Z_i} \tag{18}$$

where i denotes the pixel index, ψ is the reprojected face region which was obtained with landmarks [85], $\|\cdot\|_2$ denotes the L_2 norm, and Z_i is the occlusion attention coefficient which is described as follows.

To gain robustness to accurate texture, we set

$$Z_i = \begin{cases} 1 & \text{where the reconstructed mesh projects to} \\ 0.1 & \text{otherwise} \end{cases} \quad \text{for each pixel } i \text{ .}$$

Landmark-wise Consistency [77,86,87]. As landmarks convey the topological information of the face, we ran the faceboxes toolbox to predict 68 landmarks $\mathbf{P} \in \mathbb{R}^{68}$ as the reference. We compared the 2D landmarks of $\mathbf{I_{fina}}$ with sparse vertices of the reconstruction which correspond to these landmarks. We attained the landmarks $\mathbf{L} \in \mathbb{R}^{68}$ from the landmark vertex.

$$\mathcal{L}_{land} = \frac{1}{N} \sum_{k=0}^{N} \|\mathbf{P_k} - \mathbf{L_k}\|_2^2 \tag{19}$$

Here, $N = 68$, $\mathbf{L_k}$ denotes the 2D projection of the kth landmark vertex, and $\|\cdot\|_2$ denotes the L_2 norm.

4. Implementation Details

Our mask generation process is very similar to the traditional face parsing process. Considering the generation of the training dataset of $\mathcal{N}_{mask}$, we adopted the CelebA-HQ dataset, a high-quality version of CelebA that consists of 30,000 images at 1024×1024 resolution, each having a segmentation mask and sketch. We designed $\mathcal{N}_{mask}$ with U-Net [71] as the backbone.

To obtain $\mathbf{C_{fine}} \in \mathbb{R}^{H \times W \times 1}$, we generate contour maps using the Canny toolbox [88] as the training dataset. The sensitivity of the Canny toolbox is regulated by the standard deviation of the Gaussian smoothing filter δ. In our work, we analytically found that

$\delta \approx 1.8$ yields the best results. Our proposed network is implemented in PyTorch. We used 256×256 images with a batch size of ten to train the model of $\mathcal{N}_{\text{cont}}$. To train $\mathcal{N}_{syn}$, we followed the design of Pix2PixHD [35] with four residual blocks. The network is trained using 512×512 images with a batch size of 12. Before training the ResNet, as an initialization, we take the weights from pre-trained R-Net [77]. We set the input image size to 224×224. Our texture refinement network is designed according to the method of Lin et al. [79].

5. Experimental Results

5.1. Qualitative Comparisons with Recent Arts

Figure 3 shows our results compared with the other work. The last columns show our results. The remaining columns demonstrate the results of 3DDFA [89], PRNet [30], and DF^{2}Net [90] (Chen et al. [91]). Our results show that our results have better handled the occlusion area than other methods. Figure 3 shows that our method can reconstruct a complete face shape with geometry details under occlusion scenes, such as glasses, food and fingers. The approach of 3DDFA was aimed at extremely large poses. Therefore, it cannot reconstruct a correct face texture under occluded scenes. Other methods focused on generating high-resolution face textures rather than distinguishing occluders. At the same time, it must also be pointed out that other methods do not set up a dedicated de-occlusion component and, therefore, do not perform well under occlusion scenarios.

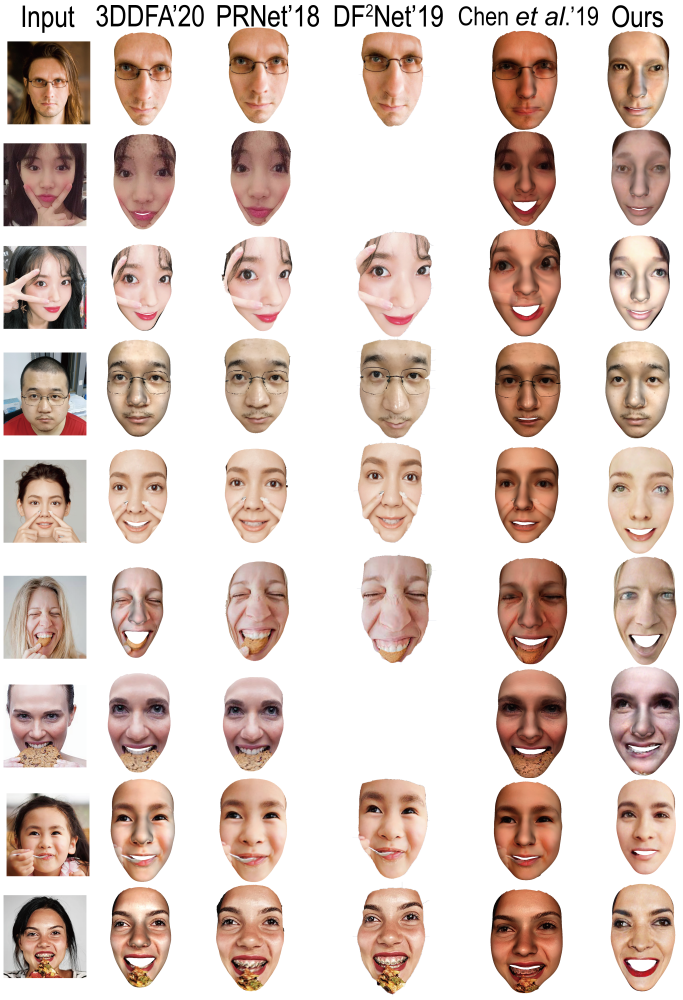

Figure 3. Comparison of qualitative results. Baseline methods from left to right: 3DDFA,PRNet, DF^{2}Net, Chen et al. and our method. The blank area means that this method does not work.

5.2. Ablation Study

In this section, we define the ablation study as a scientific examination of a deep learning system by removing its loss function blocks in order to gain insight into their effects on its overall performance. Here, we present various ablation results on the MICC and FaceWarehouse datasets [92,93]. The MICC dataset contains challenging face models of 53 subjects. For the test set, we use 90 identities with various expressions from the FaceWarehouse dataset. Table 1 shows that our ablation study produced various reconstruction evaluation results on the two datasets. Study results demonstrate that the best reconstruction results can be achieved only when the four loss functions are used fully.

Table 1. Average reconstruction errors (mm) on MICC datasets [92] and FaceWarehouse datasets [93] for ResNet trained with different loss combinations."✓" denotes employed, while "−" denotes unemployed. Our total hybrid-level loss yields considerably higher accuracy than other baselines on the two datasets.

Loss Function				MICC	Face Warehous
$\mathcal{L}_{feat}$	$\mathcal{L}_{regu}$	$\mathcal{L}_{phot}$	$\mathcal{L}_{land}$		
✓	−	−	✓	1.83 ± 0.42	2.29 ± 0.25
−	✓	✓	−	1.90 ± 0.12	1.92 ± 0.29
✓	−	✓	✓	1.88 ± 0.33	1.90 ± 0.28
−	✓	✓	−	1.78 ± 0.40	1.88 ± 0.77
✓	✓	✓	✓	1.61 ± 0.73	1.79 ± 0.57

5.3. Quantitative Comparison

5.3.1. Comparison Result on the MICC Florence Datasets

We evaluate the accuracy of the shape regression on the MICC Florence dataset [92]. The dataset is a 3D face dataset that contains 53 subjects with their ground truth 3D face scans. The ground truths are provided for 52 out of the 53 people. We artificially added some occluders (i.e., eyeglasses) as input. We calculated the average 90% largest error between the generative model and the ground truth model. Figure 4 shows that our method can effectively handle occlusion.

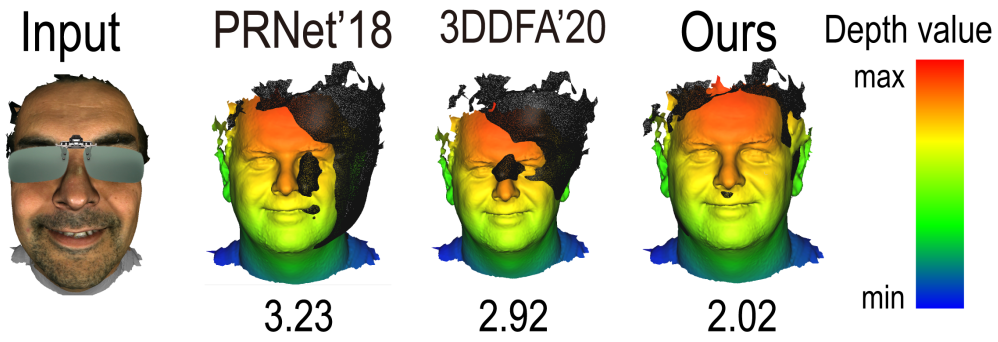

Figure 4. Comparison of error heat maps on the 3D shape recovery on MICC Florence datasets. Digits denote 90% error (mm).

5.3.2. Quantitative Comparison

The acceptance rate of face recognition is straightforward to think of as a reconstruction effect test method. Inspired by the method of Deng et al. [77], our choice of using the ResNet-50 to regress the shape coefficients ensure robustness. The basic shape is the cornerstone of our method, and we tested our approach on the Labeled Faces in the Wild datasets (LFW) [94]. Test system setting details followed the approach of Anh et al. [6].

The left of Figure 5 shows the comparison of the method of sensitivity of our method and the approach of Sela et al. It can be easily discovered that a method of Sela et al. cannot reconstruct the occluded chin. The mistake may be because their method focuses more

on local details than on the consistency of global shapes. It can be seen from Figure 5 that our method can generate a full face with the chin, which shows that this method can deal with occlusion robustly. Though 3DMM also limits the details of shape, we use it only as a foundation and add geometry details separately.

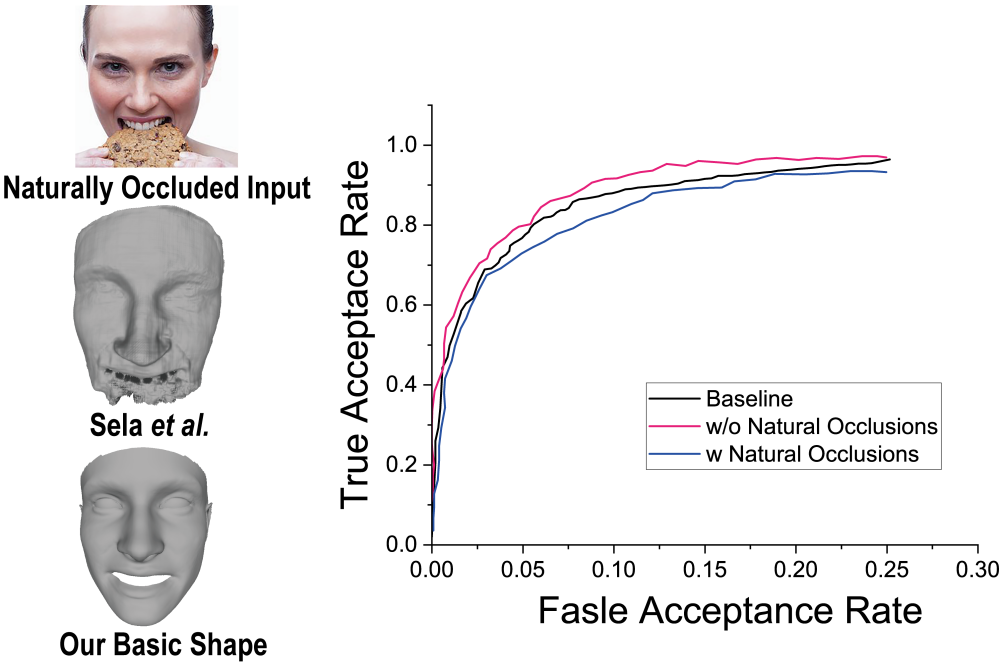

Figure 5. Basic shape reconstructions with natural occlusions. Left: Qualitative results of Sela et al. [95], and our shape. Right: LFW verification ROC for the shapes, with and without occlusions.

We further quantitatively verify the robustness of our method to occlusions. Table 2 (top) reports the verification results on the LFW benchmark [2], with and without occlusions (see also ROC in Figure 5 (right)). Though occlusion does affect the accuracy, the decline of the curve is limited, demonstrating the robustness of our approach.

Table 2. Quantitative comparison on LFW.

Method	100%-EER	Accuracy	nAUC
Tran et al.	89.40 ± 1.52	89.36 ± 1.25	95.90 ± 0.95
Our Shape and occlusions			
Ours(w/Occ)	83.89 ± 1.08	85.25 ± 0.85	89.75 ± 0.87
Ours(w/o Occ)	89.78 ± 1.21	90.33 ± 0.67	95.91 ± 0.64

6. Conclusions

In this work, we describe a method capable of producing 3D face reconstructions with convincing texture from photos taken in occluded scenes. These occlusions include fingers, food that is about to enter the mouth, glasses, and so on. At the heart of our method is its weakly supervised design, which decouples the task of estimating a robust fundamental shape from the task of estimating its mid-level details, represented here as the bump maps. Comprehensive experiments have shown that our method outperforms previous methods by a large margin in terms of both accuracy and robustness. As part of our next step, we will try to use self-supervision to reconstruct the 3D face model.

Author Contributions: Writing—review and editing of the first half of the paper, J.C. and Y.Q. ;All other work, D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is supported by Key-Area Research and Development Program of Guangdong Province (No. 2019B010150001), National Natural Science Foundation of China (No. 62072020), National Key Research and Development Program of China (No. 2017YFB1002602) and the Leading Talents in Innovation and Entrepreneurship of Qingdao (19-3-2-21-zhc).

Conflicts of Interest: We declare that we have no financial or personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "Convincing 3D face reconstruction from a single color image under occluded scenes".

References

1. Blanz, V.; Vetter, T. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1063–1074.
2. Tuan Tran, A.; Hassner, T.; Masi, I.; Medioni, G. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 5163–5172.
3. Gilani, S.Z.; Mian, A. Learning from millions of 3D scans for large-scale 3D face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1896–1905.
4. Hu, Y.; Jiang, D.; Yan, S.; Zhang, L. Automatic 3D reconstruction for face recognition. In Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, 19 May 2004; pp. 843–848.
5. Liu, X.; Chen, T. Pose-robust face recognition using geometry assisted probabilistic modeling. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 502–509.
6. Wang, S.; Cheng, Z.; Deng, X.; Chang, L.; Duan, F.; Lu, K. Leveraging 3D blendshape for facial expression recognition using CNN. *Sci. China Inf. Sci* **2020**, *63*, 120114.
7. Cao, C.; Hou, Q.; Zhou, K. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph. (TOG)* **2014**, *33*, 1–10.
8. Zhou, H.; Liu, J.; Liu, Z.; Liu, Y.; Wang, X. Rotate-and-Render: Unsupervised Photorealistic Face Rotation from Single-View Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5911–5920.
9. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep Face Recognition. 2015. Available online: https://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/parkhi15.pdf (accessed on 30 December 2021).
10. Tuan Tran, A.; Hassner, T.; Masi, I.; Paz, E.; Nirkin, Y.; Medioni, G. Extreme 3d face reconstruction: Seeing through occlusions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3935–3944.
11. Blanz, V.; Vetter, T. A morphable model for the synthesis of 3D faces. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 8–13 August 1999; Volume 99, pp. 187–194.
12. Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; Vetter, T. A 3D face model for pose and illumination invariant face recognition. In Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, Genova, Italy, 2–4 September 2009; pp. 296–301.
13. Liang, S. Data-Driven Approaches for Personalized Head Reconstruction. Ph.D. Thesis, University of Washington, Seattle, WA, USA, 2018.
14. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 681–685.
15. Saragih, J.; Goecke, R. A nonlinear discriminative approach to AAM fitting. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
16. Tzimiropoulos, G.; Pantic, M. Optimization problems for fast aam fitting in-the-wild. In Proceedings of the IEEE International Conference on Computer Vision, Cambridge, MA, USA, 20–23 June 1995; pp. 593–600.
17. Cristinacce, D.; Cootes, T.F. Feature detection and tracking with constrained local models. **2006**, *Bmvc*, *1*, 3.
18. Asthana, A.; Zafeiriou, S.; Cheng, S.; Pantic, M. Robust discriminative response map fitting with constrained local models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3444–3451.
19. Saragih, J.M.; Lucey, S.; Cohn, J.F. Deformable model fitting by regularized landmark mean-shift. *Int. J. Comput. Vis.* **2011**, *91*, 200–215.
20. Kowalski, M.; Naruniec, J.; Trzcinski, T. Deep alignment network: A convolutional neural network for robust face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 June 2017; pp. 88–97.
21. Liang, Z.; Ding, S.; Lin, L. Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. *arXiv* **2015**, arXiv: 1507.03409.

22. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 94–108.
23. Alp Guler, R.; Trigeorgis, G.; Antonakos, E.; Snape, P.; Zafeiriou, S.; Kokkinos, I. Densereg: Fully convolutional dense shape regression in-the-wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 6799–6808.
24. Yu, R.; Saito, S.; Li, H.; Ceylan, D.; Li, H. Learning dense facial correspondences in unconstrained images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017,; pp. 4723–4732.
25. Jourabloo, A.; Liu, X. Large-pose face alignment via CNN-based dense 3D model fitting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4188–4196.
26. Richardson, E.; Sela, M.; Kimmel, R. 3D face reconstruction by learning from synthetic data. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 460–469.
27. Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; Li, S.Z. Face alignment across large poses: A 3d solution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 146–155.
28. Richardson, E.; Sela, M.; Or-El, R.; Kimmel, R. Learning detailed face reconstruction from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 1259–1268.
29. Liu, F.; Zeng, D.; Zhao, Q.; Liu, X. Joint face alignment and 3D face reconstruction. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 545–560.
30. Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; Zhou, X. Joint 3d face reconstruction and dense alignment with position map regression network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 534–551.
31. Johnson, J.; Gupta, A.; Li, F.-F. Image generation from scene graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1219–1228.
32. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 1125–1134.
33. Pan, J.; Wang, C.; Jia, X.; Shao, J.; Sheng, L.; Yan, J.; Wang, X. Video generation from single semantic label map. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3733–3742.
34. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2337–2346.
35. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8798–8807.
36. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
37. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
38. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848.
39. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
40. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Salt Lake City, UT, USA, 18–23 June 2018; pp. 801–818.
41. Wei, Z.; Sun, Y.; Wang, J.; Lai, H.; Liu, S. Learning adaptive receptive fields for deep image parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, Venice, Italy, 22–29 October 2017; pp. 2434–2442.
42. Lee, C.H.; Liu, Z.; Wu, L.; Luo, P. Maskgan: Towards diverse and interactive facial image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5549–5558.
43. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 3730–3738.
44. Zhou, L.; Liu, Z.; He, X. Face parsing via a fully-convolutional continuous CRF neural network. *arXiv* **2017**, arXiv:1708.03736.
45. Yin, Z.; Yiu, V.; Hu, X.; Tang, L. End-to-end face parsing via interlinked convolutional neural networks. *Cogn. Neurodynamics* **2021**, *15*, 169–179.
46. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
47. Shen, W.; Liu, R. Learning residual images for face attribute manipulation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 4030–4038.
48. Li, M.; Zuo, W.; Zhang, D. Deep identity-aware transfer of facial attributes. *arXiv* **2016**, arXiv:1610.05586.
49. Xiao, T.; Hong, J.; Ma, J. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In Proceedings of the European Conference on Computer Vision (ECCV), Salt Lake City, UT, USA, 18–23 June 2018; pp. 168–184.

50. He, Z.; Zuo, W.; Kan, M.; Shan, S.; Chen, X. Attgan: Facial attribute editing by only changing what you want. *arXiv* **2017**, arXiv:1711.10678.
51. Lin, J.; Yang, H.; Chen, D.; Zeng, M.; Wen, F.; Yuan, L. Face Parsing with RoI Tanh-Warping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5654–5663.
52. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661.
53. Zhu, J.Y.; Krähenbühl, P.; Shechtman, E.; Efros, A.A. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 597–613.
54. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. *arXiv* **2017**, arXiv:1703.00848.
55. Demir, U.; Unal, G. Patch-based image inpainting with generative adversarial networks. *arXiv* **2018**, arXiv:1803.07422.
56. Frühstück, A.; Alhashim, I.; Wonka, P. TileGAN: Synthesis of large-scale non-homogeneous textures. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–11.
57. Li, C.; Wand, M. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 702–716.
58. Slossberg, R.; Shamai, G.; Kimmel, R. High quality facial surface and texture synthesis via generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Salt Lake City, UT, USA, 18–23 June 2018.
59. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
60. Pizzati, F.; Cerri, P.; de Charette, R. CoMoGAN: Continuous model-guided image-to-image translation. *arXiv* **2021**, arXiv:2103.06879.
61. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
62. Dapeng, Z.; Yue, Q. Generative Contour Guided Occlusions Removal 3D Face Reconstruction. In Proceedings of the 2021 International Conference on Virtual Reality and Visualization (ICVRV), Nanchang, China, 17–20 October 2021; pp. 74–79.
63. Dapeng, Z.; Yue, Q. Learning Detailed Face Reconstruction Under Occluded Scenes. In Proceedings of the 2021 International Conference on Virtual Reality and Visualization (ICVRV), Nanchang, China, 17–20 October 2021; pp. 80–84.
64. Dapeng, Z.; Yue, Q. Generative Landmarks Guided Eyeglasses Removal 3D Face Reconstruction. In *International Conference on Multimedia Modeling*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 111–122.
65. Zhao, D.; Qi, Y. Generative Face Parsing Map Guided 3D Face Reconstruction Under Occluded Scenes. In *Advances in Computer Graphics*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 252–263.
66. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
67. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
68. Dolhansky, B.; Ferrer, C.C. Eye in-painting with exemplar generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7902–7911.
69. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv* **2019**, arXiv:1901.00212.
70. Song, Y.; Yang, C.; Lin, Z.; Liu, X.; Huang, Q.; Li, H.; Kuo, C.C.J. Contextual-based image inpainting: Infer, match, and translate. In Proceedings of the European Conference on Computer Vision (ECCV), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3–19.
71. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
72. Yang, Y.; Guo, X.; Ma, J.; Ma, L.; Ling, H. LaFIn: Generative Landmark Guided Face Inpainting. *arXiv* **2019**, arXiv:1911.11394.
73. Sajjadi, M.S.; Scholkopf, B.; Hirsch, M. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–27 October 2017; pp. 4491–4500.
74. Ramamoorthi, R.; Hanrahan, P. An efficient representation for irradiance environment maps. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, New York, NY, USA, 8–13 August 2001; pp. 497–500.
75. Ramamoorthi, R.; Hanrahan, P. A signal-processing framework for inverse rendering. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, New York, NY, USA, 8–13 August 2001; pp. 117–128.
76. Müller, C. *Spherical Harmonics*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 17.
77. Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; Tong, X. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
78. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
79. Lin, J.; Yuan, Y.; Shao, T.; Zhou, K. Towards High-Fidelity 3D Face Reconstruction from In-the-Wild Images Using Graph Convolutional Networks. *arXiv* **2020**, arXiv:2003.05653.

80. Genova, K.; Cole, F.; Maschinot, A.; Sarna, A.; Vlasic, D.; Freeman, W.T. Unsupervised training for 3d morphable model regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8377–8386.
81. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
82. Tewari, A.; Zollhöfer, M.; Garrido, P.; Bernard, F.; Kim, H.; Pérez, P.; Theobalt, C. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2549–2559.
83. Tewari, A.; Zollhofer, M.; Kim, H.; Garrido, P.; Bernard, F.; Perez, P.; Theobalt, C. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1274–1283.
84. Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2387–2395.
85. Nirkin, Y.; Masi, I.; Tuan, A.T.; Hassner, T.; Medioni, G. On face segmentation, face swapping, and face perception. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 98–105.
86. Wang, X.; Guo, Y.; Deng, B.; Zhang, J. Lightweight Photometric Stereo for Facial Details Recovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 740–749.
87. Bulat, A.; Tzimiropoulos, G. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–27 October 2017; pp. 1021–1030.
88. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *6*, 679–698.
89. Guo, J.; Zhu, X.; Yang, Y.; Yang, F.; Lei, Z.; Li, S.Z. Towards Fast, Accurate and Stable 3D Dense Face Alignment. *arXiv* **2020**, arXiv:2009.09960.
90. Zeng, X.; Peng, X.; Qiao, Y. DF2Net: A Dense-Fine-Finer Network for Detailed 3D Face Reconstruction. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 2315–2324.
91. Chen, A.; Chen, Z.; Zhang, G.; Mitchell, K.; Yu, J. Photo-realistic facial details synthesis from single image. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9429–9439.
92. Bagdanov, A.D.; Del Bimbo, A.; Masi, I. The florence 2d/3d hybrid face dataset. In Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding, New York, NY, USA, 1 December 2011; pp. 79–80.
93. Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; Zhou, K. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graph.* **2013**, *20*, 413–425.
94. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Proceedings of the Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition, Marseille, France, 17 October 2008.
95. Sela, M.; Richardson, E.; Kimmel, R. Unrestricted facial geometry reconstruction using image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–27 October 2017; pp. 1576–1585.