Generative Face Parsing Map Guided 3D Face Reconstruction Under Occluded Scenes

Dapeng Zhao[1] and Yue Qi[1,2,3(✉)]

[1] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering at Beihang University, Beijing, China
qy@buaa.edu.cn
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Qingdao Research Institute of Beihang University, Qingdao, China

Abstract. Over the past few years, single-view 3D face reconstruction methods can produce beautiful 3D models. Nevertheless, the input of these works is unobstructed faces. We describe a system designed to reconstruct convincing face texture in the case of occlusion. Motivated by parsing facial features, we propose a complete face parsing map generation method guided by landmarks. We estimate the 2D face structure of the reasonable position of the occlusion area, which is used for the construction of 3D texture. An excellent anti-occlusion face reconstruction method should ensure the authenticity of the output, including the topological structure between the eyes, nose, and mouth. We extensively tested our method and its components, qualitatively demonstrating the rationality of our estimated facial structure. We conduct extensive experiments on general 3D face reconstruction tasks as concrete examples to demonstrate the method's superior regulation ability over existing methods often break down. We further provide numerous quantitative examples showing that our method advances both the quality and the robustness of 3D face reconstruction under occlusion scenes.

Keywords: 3D face reconstruction · Face parsing · Occluded scenes

1 Introduction

3D face reconstruction refers to synthesizing a 3D face model given one input face photo. It has a wide range of applications, such as face recognition and digital entertainment [25]. Existing methods mainly concentrate on unobstructed faces, thus limiting the scenarios of their actual applications. Reconstructing a 3D face model from a single photo is a classical and fundamental problem in computer vision. The reconstruction task is challenging as human face structure partial invisibility when considering occluded scenes. Over the past five years, the related problem of face inpainting in images has gradually developed to the rationality of face photo generation in the most extreme scenes [15].

N. Magnenat-Thalmann et al. (Eds.): CGI 2021, LNCS 13002, pp. 252–263, 2021.
https://doi.org/10.1007/978-3-030-89029-2_20

We cannot use artificial intelligence to robustly predict the 3D texture of the occluded area of the face. On the other hand, when faces are partially occluded, existing methods often indiscriminately reconstruct the occluded area. With the assistance of face parsing map, we find a way to identify the occluded area and reconstruct the input image to a reasonable 3D face model. The main contributions are summarized as follows:

- We propose a novel algorithm that combines feature points and face parsing map to generate face with complete facial features.
- To address the problem of invisible face area under occluded scenes, we propose synthesizing input face photo based on Generative Adversarial Network rather than reconstructing 3D face directly.
- We have improved the loss function of our 3D reconstruction framework for occluded scenes. Our method obtains state-of-the-art qualitative performance in real-world images.

2 Related Works

2.1 Generic Face Reconstruction

The classic methods use reference 3D face models to fit the input face photo. Some recent techniques use Convolution Neural Networks (CNNs) to regress landmark locations with the raw face image. Some recent techniques firstly used CNNs to predict the 3DMM parameters with input face image.

2.2 Face Image Synthesis

Deep pixel-level face generating has been studied for a few years. Many methods achieve remarkable results. EdgeConnect [12] shows impressive proceeds which disentangling generation into two stages: edge generator and image completion network. Contextual Attention [22] 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 Approach

3.1 Landmark Prediction Task

Figure 1 shows the entire process of our work.In the landmark prediction task, we found that generating accurate 68 feature points $\mathbf{Z_{lmk}} \in \mathbb{R}^{2\times 68}$ was a crucial part under occlusion scenes. The architecture $\mathcal{N}_{\mathrm{lmk}}$ aims to generate landmarks from a corrupted face photo $\mathrm{I_{cor}}$: $\mathbf{Z_{lmk}}{=}\mathcal{N}_{\mathrm{lmk}}\left(\mathrm{I_{cor}}; \theta_{lmk}\right)$, where θ_{lmk} denotes the trainable parameters. Since we want to focus more on efficiency and follow

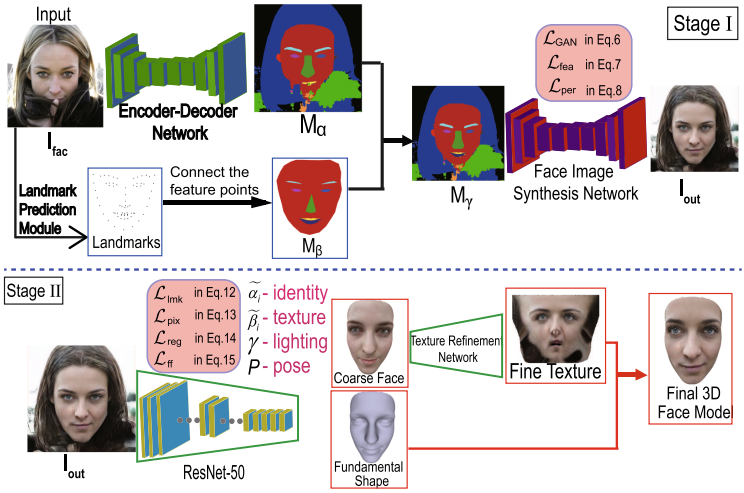

Fig. 1. Overall our pipeline. We first remove the occluded area and reconstruct the face with complete facial features. Then we utilize ResNet-50 and texture refinement network to reconstruct the final 3D model.

face parsing map generation task, we built a sufficiently effective $\mathcal{N}_{\text{lmk}}$ upon the MobileNet-V3 [6]. $\mathcal{N}_{\text{lmk}}$ is focused on feature extraction, unlike traditional landmark detectors. The final module is realized by fully connecting the fused feature maps. We set the loss function $\mathcal{L}_{lmk}$ as follows:

$$\mathcal{L}_{lmk} = \left\| \mathbf{Z}_{\mathbf{lmk}}^{(\mathbf{i})} - \hat{\mathbf{Z}}_{\mathbf{gt}}^{(\mathbf{i})} \right\|_2^2 \tag{1}$$

where $\hat{\mathbf{Z}}_{\mathbf{gt}}^{(\mathbf{i})}$ denotes the ith ground truth face landmarks.

3.2 Face Parsing Map Generation

Pixel-level recognition of occlusion and face skin areas is a prerequisite for our framework to ensure accuracy. To benefit from the annotated face dataset CelebAMask-HQ [10], we used an encoder-decoder architecture $\mathcal{N}_\alpha$ based on U-Net [17] to estimate pixel-level label classes. Given a squarely resized face image $\mathbf{I_{fac}} \in \mathbb{R}^{H\times W\times 3}$, we applied the trained face parsing model $\mathcal{N}_\alpha$ to obtain the parsing map $\mathbf{M_\alpha} \in \mathbb{R}^{H\times W\times 1}$. On the other hand, given the landmarks $\mathbf{Z_{lmk}} \in \mathbb{R}^{2\times 68}$, we connected the feature points to form a region. Then these regions can form a parsing map $\mathbf{M}_\beta \in \mathbb{R}^{H\times W\times 1}$ including facial features. Please notice that, in our work, we assumed that facial features only include only five parts, including facial skin, eyebrows, eyes, nose and lips. The final map $\mathbf{M}_\gamma \in \mathbb{R}^{H\times W\times 1}$ (see Fig. 2) without occluded objects needs $\mathbf{M_\alpha}$ plus $\mathbf{M_\beta}$. In order to generate $\mathbf{M}_\gamma$ including the complete facial features, we designed Algorithm 1.

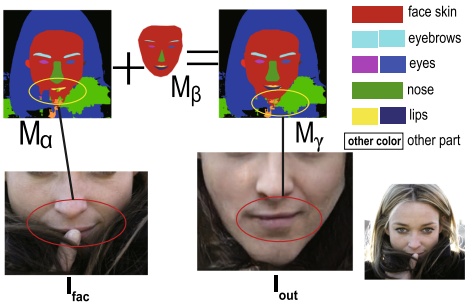

Fig. 2. Our face parsing map generation module, which follows Algorithm 1. The results shown in the figure show that our method finally successfully removed the occlusion of fingers and hair

Algorithm 1. Face Parsing Map Plus Algorithm, our proposed algorithm. All experiments in the papers Map **A** and Map **B** have the same width and height.

Input: A_i, pixels point on the face parsing map **A**. B_i, pixels point on the face parsing map **B**. $V(z)$, the function of getting the grayscale value of point z. $X(i)$, the horizontal coordinate value of i in the map. $Y(i)$, the vertical coordinate value of i in the map. W, the width of the map. H, the height of the map. S, the gray value range of the facial features area (only include four parts:eyebrows, eyes, nose, lips). O, gray value range of the facial skin area.
Input: Face parsing map **A** and **B**
Output: C_i, pixels point on the new face parsing map **C**

while $Y(i) <= \mathrm{H}$ **do** ▷ Start to generate complete face skin
  while $X(i) <= \mathrm{W}$ **do**
    if $A_i \in \mathrm{O}$ **then**
      $C_i \leftarrow A_i, X(i) + 1 \leftarrow X(i)$
    else if A_i **NOT** $\in \mathrm{O}$ **AND** $B_i \in \mathrm{O}$ **then**
      $C_i \leftarrow B_i, X(i) + 1 \leftarrow X(i)$
    else $C_i \leftarrow A_i, X(i) + 1 \leftarrow X(i)$
  end while
  $Y(i) + 1 \leftarrow Y(i)$
end while

while $Y(i) <= \mathrm{H}$ **do** ▷ Start to generate complete facial features
  while $X(i) <= \mathrm{W}$ **do**
    if $A_i \in \mathrm{S}$ **then**
      $C_i \leftarrow A_i, X(i) + 1 \leftarrow X(i)$
    else if A_i **NOT** $\in \mathrm{S}$ **AND** $B_i \in \mathrm{S}$ **then**
      $C_i \leftarrow B_i, X(i) + 1 \leftarrow X(i)$
    else $C_i \leftarrow A_i, X(i) + 1 \leftarrow X(i)$
  end while
  $Y(i) + 1 \leftarrow Y(i)$
end while

3.3 Face Image Synthesis with GAN

Face Image Synthesis Network. To benefit from the Pix2Pix architecture, we proposed a Face Image Synthesis Network (FISN) $\mathcal{N}_{\mathrm{et}}$, which was based on Pix2PixHD [26] as a backbone. FISN receives $\mathbf{I_{fac}} \in \mathbb{R}^{H\times W\times 3}$ and $\mathbf{M_{\alpha}}$ as inputs. The detailed architecture is shown in Fig. 1. To fuse $\mathbf{I_{fac}}$ and $\mathbf{M_{\alpha}}$, we used Spatial Feature Transform (SFT) layer [14] learned a mapping function $\mathcal{M}$ that outputs a parameter pair (γ, β) based on the prior condition Ψ from the features $\mathbf{M_{\alpha}}$. A pair of affine transformation parameters (γ, β) model the prior Ψ . Here, the mapping equation can be expressed as $(\gamma, \beta) = M(\Psi)$. After obtaining (γ, β) , the transformation is carried out by the SFT layer:

$$SFT(\mathbf{F_{map}}|\gamma, \beta) = \gamma \odot F + \beta \tag{2}$$

where $\mathbf{F_{map}}$ denotes the feature maps from $\mathbf{I_{fac}}$,$\odot$ denotes Hadamard product. Therefore, we conditioned spatial information $\mathbf{M_{\alpha}}$ on style data $\mathbf{I_{fac}}$ and generated affine parameters (x_i,y_i) followed (x_i,y_i)=$\mathcal{N}_{et}(I_{fac}, M_{\alpha})$. Related research [14] showed that ordinary normalization layers would "wash away" semantic information. To transfer (x_i,y_i) to new mask input $\mathbf{M_{\gamma}}$, we utilized semantic region-adaptive normalization (SEAN) [29] on residual blocks z_i in the FISN. Let H, W and C be the height, width and the number of channels in the activation map of the deep convolutional network for a batch of N samples. The modulated activation value at the site was defined as:

$$SEAN(z_i, x_i, y_i) = x_i \frac{z_i - \mu(z_i)}{\sigma(z_i)} + y_i \tag{3}$$

where $\mu(z_i)$ and $\sigma(z_i)$ are the mean and standard deviation of the activation $(n \in N, c \in C, y \in H, x \in W)$ in channel c :

$$\mu(z_i) = \frac{1}{NHW} \sum_{n,y,x} h_{n,c,y,x} \tag{4}$$

$$\sigma(z_i) = \sqrt{\frac{1}{NHW} \sum_{n,y,x} \left((h_{n,c,y,x})^2 - \mu(z_i)^2 \right)} \tag{5}$$

FISN is a generator that learns the style mapping between $\mathbf{I_{fac}}$ and $\mathbf{M_{\gamma}}$ according to the spatial information provided by $\mathbf{M_{\alpha}}$. Therefore, face features (*e.g.* eyes style) in $\mathbf{I_{fac}}$ are shifted to the corresponding position on $\mathbf{M_{\gamma}}$ so that FISN can synthesis image $\mathbf{I_{out}}$ which removed occlusion.

Loss Function. The design of our loss function for FISN is inspired by Pix2PixHD [26], MaskGAN [10] and SEAN [29], which contains three components:

(1) *Adversarial loss*. Let D_1 and D_2 be two discriminators at different scales, $\mathcal{L}_{GAN}$ is the conditional adversarial loss defined by

$$\mathcal{L}_{GAN} = \mathbb{E}\left[\log\left(\mathrm{D}_{1,2}\left(\mathbf{I_{fac}}, \mathbf{M_{\alpha}}\right)\right)\right] + \mathbb{E}\left[1 - \log\left(\mathrm{D}_{1,2}\left(\mathbf{I_{out}}, \mathbf{M_{\alpha}}\right)\right)\right] \tag{6}$$

(2) *Feature matching loss* [26]. Let T be the total number of layers in discriminator D .$\mathcal{L}_{fea}$ is the feature matching loss which computed the L_1 distance between the real and generated face image defined by

$$\mathcal{L}_{\text{fea}}=\mathbb{E}\sum_{i=1}^{T}\left\|D_{1,2}^{(i)}\left(\mathbf{I_{fac}},\mathbf{M}_{\boldsymbol{\alpha}}\right)-D_{1,2}^{(i)}\left(\mathbf{I_{out}},\mathbf{M}_{\alpha}\right)\right\|_1 \tag{7}$$

(3) *Perceptual loss* [8]. Let N be the total number of layers used to calculate the perceptual loss, $F^{(i)}$ be the output feature maps of the ith layer of the VGG network [21].$\mathcal{L}_{per}$ is the perceptual loss which computes the L_1 distance between the real and generated face image defined by

$$\mathcal{L}_{\text{per}}=\mathbb{E}\sum_{i=1}^{N}\frac{1}{M_i}[\left\|F^{(i)}\left(\mathbf{I_{fac}}\right)-F^{(i)}\left(\mathbf{I_{out}}\right)\right\|_1] \tag{8}$$

The final loss function of FISN used in our experiment is made up of the above-mentioned three loss terms as:

$$\mathcal{L}_{FISN}=\mathcal{L}_{GAN}+\lambda_1\mathcal{L}_{\text{fea}}+\lambda_2\mathcal{L}_{\text{per}} \tag{9}$$

where we set λ_1=λ_2= 10 respectively in our experiments.

3.4 Camera and Illumination Model

Given an face image, we adopt the Basel Face Model (BFM) [16]. After the 3D face is reconstructed, it can be projected onto the image plane with the perspective projection:

$$V_{2d}\left(\mathbf{P}\right)=f*\mathbf{P_r}*\mathbf{R}*\mathbf{S_{mod}}+\mathbf{t_{2d}} \tag{10}$$

where $V_{2d}\left(\mathbf{P}\right)$ 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, $\mathbf{S_{mod}}$ denotes the shape of the face and $\mathbf{t_{2d}}\in\mathbb{R}^3$ denotes the translation vector.

We approximated the scene illumination with Spherical Harmonics (SH) [3] for face. Thus, we can compute the face as Lambertian surface and skin texture follows:

$$\mathbf{C}\left(\mathbf{r_i},\mathbf{n_i},\boldsymbol{\gamma}\right)=\mathbf{r_i}\odot\sum_{b=1}^{B^2}\boldsymbol{\gamma_b}\Phi_b\left(\mathbf{n_i}\right) \tag{11}$$

where $\mathbf{r_i}$ denotes skin reflectance, $\mathbf{n_i}$ denotes surface normal, $\odot$ denotes the Hadamard product, $\boldsymbol{\gamma}\in\mathbb{R}^9$ under monochromatic lights condition, $\Phi_b:\mathbb{R}^3\rightarrow\mathbb{R}$ denotes SH basis function, B denotes the number of spherical harmonics bands and $\boldsymbol{\gamma_b}\in\mathbb{R}^3$ (here we set $B=3$) denotes the corresponding SH coefficients.

Therefore, parameters to be learned can be denoted by a vector $\boldsymbol{y}=(\widetilde{\boldsymbol{\alpha_i}},\widetilde{\boldsymbol{\beta_i}},\boldsymbol{\gamma},\boldsymbol{p})\in\mathbb{R}^{175}$, where $\mathbf{p}\in\mathbb{R}^6=\{\boldsymbol{pitch},\boldsymbol{yaw},\boldsymbol{roll},\boldsymbol{f},\boldsymbol{t_{2D}}\}$ denotes face

poses. In this work, we used a fixed ResNet-50 [5] network to regress these coefficients. The loss function of ResNet-50 follows Eq. 16. We then got the fundamental shape $\mathbf{S_{base}}$ (coordinate,*e.g.* x, y, z) and the coarse texture $\mathbf{T_{coa}}$ (albedo,*e.g.* r, g, b). We used a coarse-to-fine network based on graph convolutional networks of Lin *et al.* [11] for producing the fine texture $\mathbf{T_{fin}}$.

3.5 Loss Function of 3D Reconstruction

Given a generated image $\mathbf{I_{out}}$,we used the ResNet to regress the corresponding coefficient y. The design of loss function for ResNet contained four components:

(1) Landmark Loss. As facial landmarks convey the structural information of the human face, we used landmark loss to measure how close projected shape landmark vertices to the corresponding landmarks in the image $\mathbf{I_{out}}$. We ran the landmark prediction module $\mathcal{N}_{lmk}$ to detect 68 landmarks $\left\{z_{lmk}^{(n)}\right\}$ from the training images. We obtained landmarks $\left\{l_y^{(n)}\right\}$ from rendering facial images. Then, we computed the loss as:

$$\mathcal{L}_{lmk}(y) = \frac{1}{N}\sum_{\mathrm{n}=1}^{N}\left\|\mathrm{z}_{\mathrm{lmk}}^{(n)} - l_y^{(n)}\right\|_2^2 \tag{12}$$

where $\|\cdot\|_2$ denotes the L_2 norm.

(2) Accurate Pixel-wise Loss. The rendering layer renders back an image $\mathbf{I}_{\mathbf{y}}^{(i)}$ to compare with the image $\mathbf{I}_{\mathbf{out}}^{(i)}$. The pixel-wise loss is formulated as:

$$\mathcal{L}_{\mathrm{pix}}(y) = \frac{\sum_{i\in\mathcal{M}} P_i \cdot \left\|\mathbf{I}_{\mathbf{out}}^{(i)} - \mathbf{I}_{\mathbf{y}}^{(i)}\right\|_2}{\sum_{i\in\mathcal{M}} P_i} \tag{13}$$

where i denotes pixel index, $\mathcal{M}$ is the reprojected face region which obtained with landmarks [13], $\|\cdot\|_2$ denotes the L_2 norm and P_i is occlusion attention coefficient which is described as follows. To gain robustness to accurate texture, we set $P_i = \begin{cases} 1 & \text{if } i \in \text{facial features of } M_\alpha \\ 0.1 & \text{otherwise} \end{cases}$ for each pixel i.

(3) Regularization Loss. To prevent shape deformation and texture degeneration, we introduce the prior distribution to the parameters of the face model. We add the regularization loss as:

$$\mathcal{L}_{\mathrm{reg}} = \omega_\alpha \|\widetilde{\boldsymbol{\alpha}_{\mathbf{i}}}\|^2 + \omega_\beta \left\|\widetilde{\boldsymbol{\beta}_{\mathbf{i}}}\right\|^2 \tag{14}$$

here, we set ω_α=1.0, ω_β = 1.75e-3 respectively.

(4) Face Features Level Loss. To reduce the difference between 3D face with 2D image, we define the loss at face recognition level. The loss computes

the feature difference between the input image $\mathbf{I_{out}}$ and rendered image $\mathbf{I_y}$. We define the loss as a cosine distance:

$$\mathcal{L}_{ff}=1-\frac{< G(\mathbf{I_{out}}), G(\mathbf{I_y}) >}{\|G(\mathbf{I_{out}})\| \cdot \|G(\mathbf{I_y})\|} \tag{15}$$

where $G(\cdot)$ denotes the feature extraction function by FaceNet [19], $< \cdot, \cdot >$ denotes the inner product.

In summary, the final loss function of 3D face reconstruction used in our experiment is made up of the above-mentioned four loss terms as:

$$\mathcal{L}_{3D}=\lambda_3\mathcal{L}_{lmk} + \lambda_4\mathcal{L}_{pix} + \lambda_5\mathcal{L}_{reg} + \lambda_6\mathcal{L}_{ff} \tag{16}$$

where we set $\lambda_3= 1.6e-3, \lambda_4= 1.4, \lambda_5=$ 3.7e-4, $\lambda_6= 0.2$ respectively in all our experiments.

4 Implementation Details

Considering the question of landmark predictor, the 300-W dataset [18] has labeled ground truth landmarks, while the CelebA-HQ dataset [9] does not. We generated the ground truth of CelebA-HQ by the Faceboxes predictor [28] as the reference. In experiments shown in this work, we use the 256×256 images for training the landmark predictor $\mathcal{N}_{lmk}$ and the batch size = 16 . The learning rate of $\mathcal{N}_{lmk}$ is $10e-4$. We use the trained face parsing model $\mathcal{N}_{\alpha}$ [10] to generate $\mathbf{M_{\alpha}}$. We obtain $\mathbf{M}_{\gamma}$ according to Algorithm 1. FISN follows the design of Pix2PixHD [26] with four residual blocks. To train the FISN, we used the CelebAMask-HQ dataset which has 30000 semantic labels with a size of 512×512. Each label clearly marked the facial features of the face.

FISN does not use any ordinary normalization layers (*e.g.* Instance Normalization) which will wash away style information. Before training the ResNet, we take the weights from pre-trained of R-Net [3] as initialization. We set the input image size to 224×224 and the number of vertices to 35709. We design our texture refinement network based on the Graph Convolutional Network method of Lin *et al.* [11]. We do not adopt any fully-connected layers or convolutional layers in the refinement network refer to related research [11]. This will reduce the performance of the module.

5 Experimental Results

5.1 Qualitative Comparisons with Recent Works

Figure 3 shows our results compared with the other work. The last two columns show our results.The remaining columns demonstrate the results of 3DDFA [4], DF^2Net [27] and Chen *et al.* [2]. Qualitative results show that our method surpasses other methods. Figure 3 shows that our method can reconstruct a complete face model under occlusion scenes such as glasses, jewelry, palms, and hair. Other methods focused on generating high-resolution face textures. These frameworks cannot effectively deal with occluded scenes.

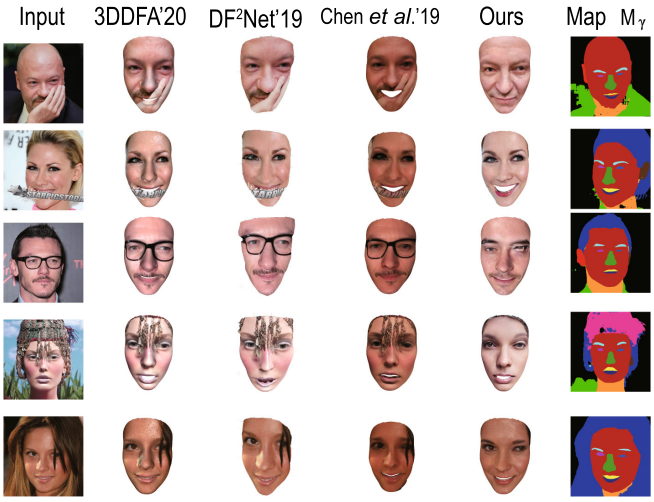

Fig. 3. Comparison of qualitative results. Baseline methods from left to right: 3DDFA, DF^{2}Net, Chen *et al.* and our method.

5.2 Quantitative Comparison

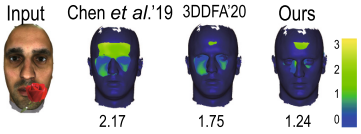

Fig. 4. Comparison of error heat maps on the MICC Florence datasets. Digits denote 90% error (mm).

Comparison Result on the MICC Florence Datasets. MICC Florence dataset [1] is a 3D face dataset that contains 53 faces with their ground truth models. We artificially added some occluders 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.

Occlusion Invariance of the Foundation Shape. Our choice of using the ResNet-50 to regress the shape coefficients is motivated by the unique robustness to extreme viewing conditions in the paper of Deng *et al.* [3]. To fully support the application of our method to occluded face images, we test our system on the Labeled Faces in the Wild datasets (LFW) [7]. We used the same face test system from Anh *et al.* [24], and we refer to that paper for more details.

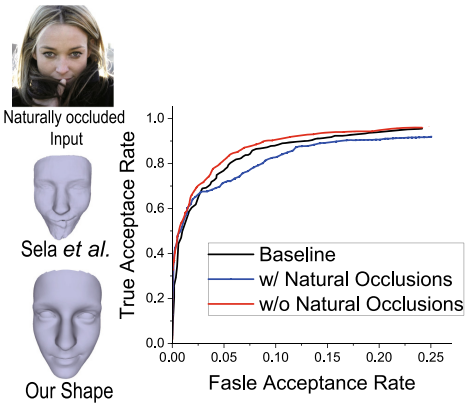

Fig. 5. Reconstructions with occlusions. Left: Qualitative results of Sela *et al.* [20] and our shape. Right: LFW verification ROC for the shapes, with and without occlusions.

Figure 5 (left) shows the sensitivity of the method of Sela et al. [20]. Their result clearly shows the outline of a finger. Their failure may be due to more focus on local details, which weakly regularizes the global shape. However, our method recognizes and regenerates the occluded area. Our method much robust provides a natural face shape under common occlusion scenes. Though 3DMM also limits the details of shape, we use it only as a foundation and add refined texture separately.

Table 1. Quantitative evaluations on LFW.

Method	100%-EER	Accuracy	nAUC
Tran *et al.* [23]	89.40 ± 1.52	89.36 ± 1.25	95.90 ± 0.95
Ours (w/ Occ)	85.75 ± 1.12	86.49 ± 0.97	93.89 ± 1.31
Ours (w/o Occ)	90.57 ± 1.43	89.87 ± 0.71	96.59 ± 0.37

We further quantitatively verify the robustness of our method to occlusions. Table 1 (top) reports verification results on the LFW benchmark with and without occlusions (see also ROC in Fig. 5 (right)). Though occlusions clearly impact recognition, this drop of the curve is limited, demonstrating the robustness of our method.

6 Conclusions

In this work, we present a novel single-image 3D face reconstruction method under occluded scenes with high fidelity textures. Comprehensive experiments have shown that our method outperforms previous methods by a large margin

in terms of both accuracy and robustness. Future work includes combining our method with Transformer architecture to further improve accuracy.

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

References

1. 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, pp. 79–80 (2011)
2. 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, pp. 9429–9439 (2019)
3. 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 (2019)
4. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. arXiv preprint arXiv:2009.09960 (2020)
5. 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, pp. 770–778 (2016)
6. Howard, A., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)
7. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition (2008)
8. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
9. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
10. 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, pp. 5549–5558 (2020)
11. Lin, J., Yuan, Y., Shao, T., Zhou, K.: Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. arXiv preprint arXiv:2003.05653 (2020)
12. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019)
13. Nirkin, Y., Masi, I., Tuan, A.T., Hassner, T., Medioni, G.: On face segmentation, face swapping, and face perception. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 98–105. IEEE (2018)

14. 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, pp. 2337–2346 (2019)
15. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
16. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301. IEEE (2009)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
18. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 397–403 (2013)
19. 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, pp. 815–823 (2015)
20. 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, pp. 1576–1585 (2017)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Song, Y., et al.: Contextual-based image inpainting: Infer, match, and translate. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
23. 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, pp. 5163–5172 (2017)
24. Tun Trn, 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, pp. 3935–3944 (2018)
25. 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 **63**(120114), 1–120114 (2020)
26. 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, pp. 8798–8807 (2018)
27. 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, pp. 2315–2324 (2019)
28. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: Faceboxes: a cpu real-time face detector with high accuracy. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–9. IEEE (2017)
29. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5104–5113 (2020)