3D Face Reconstruction with Geometry Details from a Single Color Image 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. 3D face reconstruction technology aims to generate a face stereo model naturally and realistically. Previous deep face reconstruction approaches are typically designed to generate convincing textures and cannot generalize well to multiple occluded scenarios simultaneously. By introducing bump mapping, we successfully added mid-level details to coarse 3D faces. More innovatively, our method takes into account occlusion scenarios. Thus on top of common 3D face reconstruction approaches, we in this paper propose a unified framework to handle multiple types of obstruction simultaneously (e.g., hair, palms and glasses *et al.*). Extensive experiments and comparisons demonstrate that our method can generate high-quality reconstruction results with geometry details from captured facial images under occluded scenes.

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

1 Introduction

High-quality 3D face reconstruction is a fundamental problem in computer graphics [31] that is related to various applications such as digital animation [32], video editing [32] and face recognition [40,41]. Since Vetoer's first 3D face [35], 3D reconstruction methods have rapidly advanced enabling applications. However, these methods all perform poorly in terms of face geometry details. To make the problem tractable, most proposed methods introduce existing statistical models or prior knowledge. These models are unable to reconstruct expression-dependent wrinkles, which are essential for analyzing human expression.

Several methods recover detailed facial geometry that lacks robustness to occlusions [1,9]. We introduce a novel face geometry detail generation method, which learns bump maps (simulate geometry changes) from in-the-wild face images with occlusion. In contrast to prior work (estimating mid-level features often breaks down), our method generates bump maps from a low-dimensional representation containing subject-specific detail parameters and

E. Pimenidis et al. (Eds.): ICANN 2022, LNCS 13530, pp. 332–344, 2022.
https://doi.org/10.1007/978-3-031-15931-2_28

expression parameters. Our detailed model builds upon this separation design. This design is fundamental, as it allows estimating a robust global shape, even under occluded scenes.

The main contributions are summarized as follows:

- We propose a novel Face Image Synthesis Network, a simple yet effective diversity promoting face image regeneration approach. The regenerated eyeglasses removal face without glasses will guide the generation of a 3D model.
- We have improved the loss function of our 3D reconstruction system for occluded scenes with eyeglasses. Our results are more accurate than other approaches. As a result of our method, we are able to obtain state-of-the-art qualitative performance in real-world images.

2 Related Work

2.1 Single Image 3D Face Reconstruction

Since the first 3DMM model was proposed by Blanz and Vetter [2], single image based 3D face reconstruction has become a hot research topic and considerable progress have been made in the field. Richardson *et al.* [25] presented a method based on CNN that can reconstruct 3D face based on synthetic data. As training deep neural networks usually demand a large amount of data to get acceptable results, Deng *et al.* [5] proposed an approach that can achieve accurate 3D face reconstruction with weakly supervised learning based on less training data. Kemelmacher-Shlizerman and Basri [11] recovered 3D faces by exploiting the similarity of faces based on a single 3D reference model of a different person. Liu *et al.* [19] built a 3D face model that can exploit both faces with fully labeled 3D landmarks and unlimited unlabeled in-the-wild face images. Lee *et al.* [16] employed an uncertainty-aware encoder and a fully nonlinear decoder model for realistic 3D face reconstruction. Cheng *et al.* [3] solved the 3D face reconstruction problem based on graph convolutional networks obtaining good results without scarifying speed. Shang *et al.* [30] proposed a self-supervised training architecture that is accurate and robust, even under large variations of expressions, poses, and illumination conditions. Li *et al.* [17] publicized an end-to-end framework and designed an efficient network model that can apparently increase the accuracy of face alignment and 3D face reconstruction. Li *et al.* [18] presented a multi-attribute regression reconstruction network that can work well in complex cases when provided with 2D images including severe poses, extreme expressions, and partial occlusions.

2.2 Generative Adversarial Networks

Generative adversarial networks (GANs) was first proposed by Goodfellow *et al.* to study the generative model. Classical GANs consist of a generator and a discriminator. The aim of the generator is to generate data samples that can confuse the discriminator. The generator and the discriminator must improve themselves

to win the 'game' until a Nash equilibrium is achieved; then generator successfully learns the distribution of the real dataset. GANs have been applied in many fields, including face image synthesis. Zhan *et al.* [39] proposed Spatial Fusion GAN (SF-GAN), which can obtain better results in both geometry and appearance spaces utilizing a geometry synthesizer and an appearance synthesizer. A triple-translation GAN (TTGAN) is proposed for face image synthesis by Ye *et al.* [38]. TTGAN adopts a triple translation consistency loss to translate from a rendered original input image to the desired output image. Sangloy *et al.* [28] proposed an adversarial image synthesis architecture that can extract information from sketched boundaries and parse color strokes and output realistic face images.

2.3 Face Image Synthesis

Deep pixel-level face generating has been studied for several years. Many methods achieve remarkable results. Context encoder [23] is the first deep learning network designed for image inpainting with encoder-decoder architecture. Nevertheless, the networks do a poor job in dealing with human faces. Following this work, Yang *et al.* used a modified VGG network to improve the result of the context-encoder, by minimizing the feature difference of photo background. Dolhansky *et al.* 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 [21] shows impressive proceeds which disentangling generation into two stages: edge generator and image completion network. Contextual Attention 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 Proposed Approach

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

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

3.1 Face Parsing Map Generation

Our goal is to realize detailed 3D face shape reconstruction under occluded scenes using our method. Pixel-level recognition of eyeglasses areas serves as a key

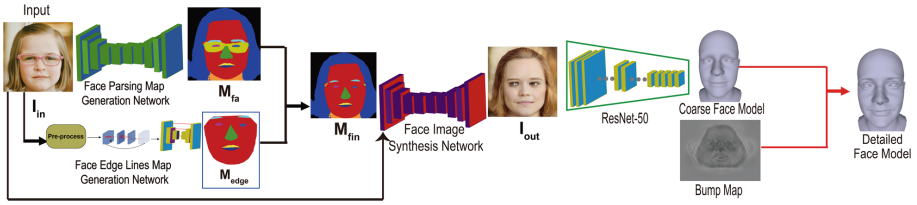

Fig. 1. Method overview. At first, as input for our face image synthesis network, we need the target image $\mathbf{I_{in}}$ and map $\mathbf{M_{fin}}$. We utilize the face parsing map generation module and edge lines map generation module to obtain the map $\mathbf{M_{fa}}$ and $\mathbf{M_{edge}}$. Then we obtain the final face parsing map $\mathbf{M_{fin}}$ following Zhao *et al.*'s Algorithm [42]. After obtaining the face image $\mathbf{I_{out}}$ with eyeglasses removed, in step two, we leverage ResNet-50 and texture refinement network to reconstruct the final 3D model.

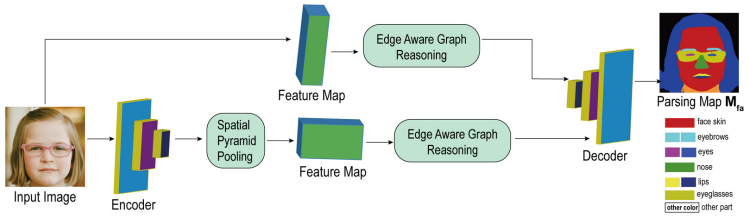

Fig. 2. The overview of the proposed face parsing network.

step for our framework to ensure accuracy. Face parsing is a fundamental facial analysis task. Recently, methods based on Fully Convolutional Networks have achieved remarkable results on this task [8,20,36]. As shown in Fig. 2, given a squarely resized face image $\mathbf{I_{in}} \in \mathbb{R}^{H \times W \times 3}$, we aim to apply a modified encoder-decoder network $\mathcal{N}_{fa}$ as the backbone frame for face parsing. We take $\mathcal{N}_{fa}$ to extract features at different levels for multi-scale illustration. In the structure of $\mathcal{N}_{fa}$, high-level features contain semantic information while low-level features show local details, both of which are essential for face parsing. We feed the feature map with multi-scale information into the Edge Aware Graph Reasoning module, targeting to learn fundamental graph illustration for the characterization of the relations between vertices. The reasoning module consists of three components: graph projection operation, graph reasoning operation and graph reprojection operation. Let us make it clear. The graph projection operation projects the initial information onto vertices. The graph reasoning operation reasons the relational expression between regions over the graph and projects the acquired graph interpretation back to previous pixel grids. The graph reprojection operation leads to an optimized feature map with the same dimension and size. We implemented the reasoning module following the method of Gusi *et al.* [31]. Let us explain the last step of the network. We transmit the optimized features into a decoder to estimate the final pixel labels. In our network, two different level feature maps are concatenated into the decoder. The two feature maps are concatenated by the 1×1 convolution layer. The specific fusion method

is through upsampling. That is, the high-level feature map is upsampled to the same dimension as the low-level feature map. Finally, we obtain the face parsing map $\mathbf{M_{fa}} \in \mathbb{R}^{H \times W \times 1}$.

3.2 Face Edge Lines Map Generation

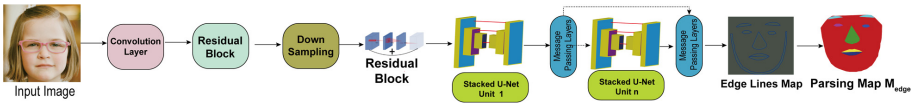

Fig. 3. The overview of the proposed face edge lines map generation approach.

In order to generate an accurate face parsing map, our method uses face edge lines to guide the reconstruction of the face parsing map. Face edge lines is closely related to the facial landmark. The reason why we choose face edge lines instead of landmarks is that landmarks have difficulties in presenting the accurate facial features structure [37]. In this section, we describe the proposed face edge lines map generation framework in detail. As shown in Fig. 3 (a) and (b), the proposed framework consists of two parts: (a) face edge lines generation module; (b) face edge lines effectiveness discriminator.

As shown in Fig. 3 (a), stacked U-Nets is the core part of the face edge lines generation module. More than piecemeal landmarks, face edge lines can well describe the geometry structure of a face. Most of the previous convolutional networks only use the convolutional features of the last layer. Image information at other scales will be lost. Unlike the previous network, the main contribution of the stacked U-Nets unit [22,26] is to use multi-scale features to represent image information. We Leverage the mean squared error (MSE) between the estimated Face edge lines map and the ground-truth map. The presence of obstructions (this paper focuses on eyeglasses) will significantly affect the accuracy of edge lines generation. In order to relieve the loss of image information due to eyeglasses, we introduce message passing layers to pass information between face edge lines. It is proposed in this implementation that the feature map at the end of each stack should be divided into M (the number) areas. We implemented the message passing approach following the method of Chu *et al.* [4]. This process is visualized in Fig. 3.

Intra-level Message Passing Layer. Among the steps involved in dealing with the problem of occlusion of eyeglasses, the intra-level message passing plays a crucial role. A layer such as this one is used at the end of each U-Nets stack in order to transmit information between visible edge lines and eyeglasses areas. Consequently, in the process of designing eyeglasses, the prediction of the eyeglasses areas can be improved through the visible edge lines data.

Inter-level Message Passing Layer. It is true that there are various U-Nets stacks that focus on different dimensions of facial information, but in the case

of multiple stacks, the facial information is transferred in the different stacks by performing communication between the former stacks and the latter stacks. When stacking more hourglass subnets, inter-level message passing is adopted to ensure that the face edge lines map maintains the quality when messages are passed from the lower stacks to the higher stacks.

Adversarial Learning for Edge Lines Effectiveness. Poor face edge lines map will adversely affect the accuracy of the 3D face model. When training, we use adversarial learning between the estimated edge lines map and the ground-truth map in order to guarantee the effectiveness of the edge lines map obtained in the generation stage. Using the Face edge lines map generator, the edge lines map $\mathbf{M_{edge}} \in \mathbb{R}^{H\times W\times 1}$ is generated with the coordinate set S_{coor}; the mapping between the generated coordinate set and the ground-truth distance matrix $\mathbf{MA_{gt}}$. In order to determine whether a generated edge line map is fake or not, the ground truth d_{gt} can be calculated as:

$$d_{gt}(\mathbf{M_{edge}}, S_{coor}) = \begin{cases} 0, Est_{s\in S_{coor}}(d_{gt} < \theta) < \delta \\ 1, \text{other cases} \end{cases} \tag{1}$$

where Est denotes the probability value calculation function, θ denotes the distance threshold to ground truth edge lines, δ denotes the probability threshold.

In order to combine the edge lines effectiveness discriminator D and the face edge lines map estimator G, we apply the concept of adversarial learning. The loss function of the discriminator D can be calculated as:

$$\mathcal{L}_D = \mathbb{E}[\log(1 - |D(G(\mathbf{I_{in}})) - d_{gt}|)] - \mathbb{E}[\log D(\mathbf{M_{gt}})] \tag{2}$$

where $\mathbf{M_{gt}}$ denotes the ground truth face edge lines map. A discriminator is trained to predict an edge lines map on the ground truth as well as predict the generated edge lines map according to d_{gt} . With effectiveness discriminator, the adversarial loss can be calculated as:

$$\mathcal{L}_{adv-loss} = \mathbb{E}\left[\log(1 - D(G(\mathbf{I_{in}}))\right] \tag{3}$$

3.3 Recovering 3D Face Geometric Details

We obtain the final face parsing map $\mathbf{M_{fin}}$ following Zhao *et al.*'s Algorithm [42]. We synthesize the face photo $\mathbf{I_{out}}$ by existing methods [15]. Given $\mathbf{I_{out}}$, we used the ResNet to regress the corresponding coefficient y. Due to the collection of large scale high-resolution 3D texture datasets is still very costly and scarce, the ResNet was trained under weakly supervised. The corresponding loss function consists of four parts [2,5]:

$$\mathcal{L}_{shape} = \lambda_{feat}\mathcal{L}_{feat} + \lambda_{regu}\mathcal{L}_{regu} + \lambda_{phot}\mathcal{L}_{phot} + \lambda_{land}\mathcal{L}_{land} \tag{4}$$

Here we set λ_{feat}= 0.2,λ_{regu} = 3.6$e-$4,λ_{phot} = 1.4, λ_{land} = 1.6$e-$3 respectively in all our experiments.

The addition of human face geometric details is the core of our method. We choose to add a bump map on the base shape $\mathbf{S_{basi}}$. Inspired by the method of image-to-image translation method, we define the displacements of the depth map as the distances through the pixels of $\mathbf{I_{out}}$ to the 3D face surface. Generally, we define the bump map $\Phi(\mathbf{b})$ as:

$$\Phi(\mathbf{b})=\begin{cases}\phi(0) \quad \text{othercases}\\ \phi(d'(\mathbf{b})-d(\mathbf{b})) \text{ face projects to } \mathbf{b}\end{cases} \tag{5}$$

where $\phi(\cdot)$ denotes an encoding function that converts the depth value to the linear range $[0,\ldots,255]$, $\mathbf{b}$ denotes the pixel coordinate $[x,y]$ in $\mathbf{I_{out}}$, $d'(\mathbf{b})$ denotes the depth, which is the distance from the surface of the detailed face shape to $\mathbf{b}$ along the line of sight, $d(\mathbf{b})$ denotes the depth of the basic shape.

Thus, Given a bump map Φ and the depth of the basic shape,we can compute the detailed depth follows $d'(\mathbf{b})=d(\mathbf{b})+\phi^{-1}(\Phi(\mathbf{b}))$. In order to increase geometric details and to suppress noise, we define the loss function as follows:

$$\mathcal{L}_{geo}=\left\|\tilde{\Phi}-\Phi\right\|+\left\|\frac{\partial\tilde{\Phi}}{\partial x}-\frac{\partial\Phi}{\partial x}\right\|+\left\|\frac{\partial\tilde{\Phi}}{\partial y}-\frac{\partial\Phi}{\partial y}\right\| \tag{6}$$

where $\|\cdot\|$ denotes the L_1 norm, $\tilde{\Phi}$ denotes the ground truth and $\frac{\partial\tilde{\Phi}}{\partial x}$, $\frac{\partial\tilde{\Phi}}{\partial y}$ denotes the 2D gradient of the bump map. After the 3D face is reconstructed, it can be projected onto the image plane with the perspective projection:

$$V_{2d}\left(\mathbf{P}\right)=f*\mathbf{P_r}*\mathbf{R}*\mathbf{S_{mod}}+\mathbf{t_{2d}} \tag{7}$$

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 and $\mathbf{t_{2d}}\in\mathbb{R}^3$ denotes the translation vector.

Therefore, we approximated the scene illumination with Spherical Harmonics (SH) [24] parameterized by coefficient vector $\gamma\in\mathbb{R}^9$. In summary, the unknown parameters to be learned can be denoted by a vector $y=(\boldsymbol{\alpha_{id}},\boldsymbol{\beta_{exp}},\boldsymbol{\beta_t},\boldsymbol{\gamma},\mathbf{p})\in\mathbb{R}^{239}$, where $\mathbf{p}\in\mathbb{R}^6=\{\mathbf{pitch},\mathbf{yaw},\mathbf{roll},f,\mathbf{t_{2D}}\}$ denotes face poses. In this work, we used a fixed ResNet-50 network to regress these coefficients.

We found that by adding these last two terms of loss function and we reduce bump map noise by favoring smoother surfaces. At the same time, the final effect shows that high-frequency details are preserved.

4 Implementation Details

All the networks were trained using the Adam solver [12]. To train our face parsing map generation network, we collected two sources dataset: Helen dataset [14] and CelebAMask-HQ dataset [15]. The Helen dataset contains 2330 images with 11 categories: background, skin, paired lips, paired eyes, paired brows,

paired mouth and hair. The CelebAMask-HQ dataset is a large-scale face parsing datasets which includes 30000 high-resolution portrait images. The dataset contains 19 categories. In addition to the facial unit, the components such as eyeglass, earring, necklace, neck, and cloth are also annotated.

In the face parsing map generation stage, our backbone is a modified version of the trained parsing model [31]. We made the parsing model exclude the average pooling layer. For the pyramid pooling module, we follow the implementation of the method of Te *et al.* [31] with exploiting global contextual information. We leveraged the fixed parsing model to generate $\mathbf{M_{fa}}$. In the face edge lines map generation stage, all training images are cropped and resized to 512×512. We obtained $\mathbf{M_{edge}}$ according the lines map generation network. We implemented message passing module following naturally obtains face features in different sizes. In the above two stages, we train our network on four datasets including 300W (3148 sample images) [27] and AFLW (24386 sample images) [13].

5 Experimental Results

In this work, we aim to generate a wide range of diverse and yet realistic 3D detailed reconstructions from occluded face images. Our approach should be characterized by the following three qualities: 1) the reconstructed geometry should fit as convincingly as possible to the visible regions, 2) the reconstructed model texture should not include eyeglasses, which is the essential requirement for the accuracy of the reconstruction.

5.1 Qualitative Comparisons with Recent Art

Figure 4 shows our results compared with the other arts. The last columns show our results. The remaining columns demonstrate the results of Sela *et al.* [29], PRNet [6] and 3DDFA [7]. Our results show that our results have better handled the occlusion area than other methods. Figure 4 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 detailed face model under occluded scenes. Its shape lacks details. Other methods focused on generating high-resolution face textures instead of geometry details. At the same time, it must also be pointed out, the other methods cannot effectively deal with occluded scenes.

5.2 Quantitative Comparison with Recent Art

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.* [5]. 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) [10]. We used the same face test system from Anh *et al.* [34], and we refer to that paper for more details.

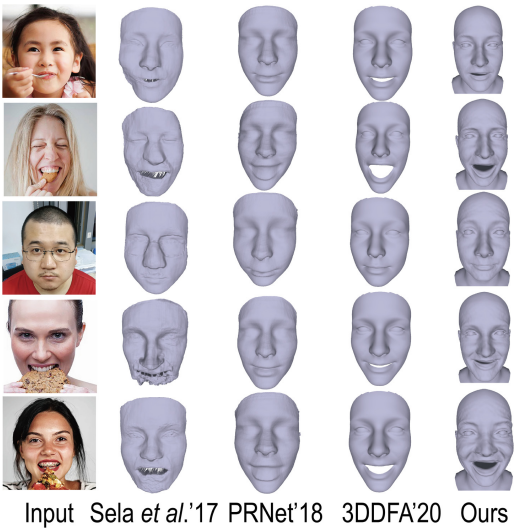

Fig. 4. Comparison of qualitative results. Baseline methods from left to right: Sela *et al.*, PRNet, 3DDFA and our method.

Figure 5 (left) shows the sensitivity of the method of Sela *et al.* [29]. Their result clearly shows the outline of the eyeglasses. 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 eyeglasses scenes. Though 3DMM also limits the details of shape, we use it only as a foundation and add refined texture separately.

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

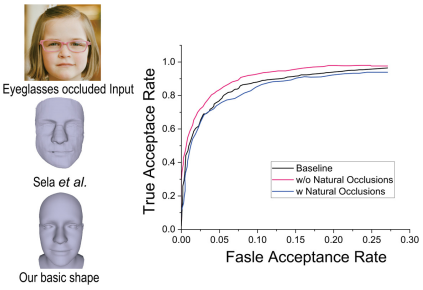

Fig. 5. Reconstructions with eyeglasses. Left: qualitative results of Sela *et al.* [29] and our shape. Right: LFW verification ROC for the shapes, with and without eyeglasses.

Table 1. Quantitative evaluations on LFW.

Method	100%-EER	Accuracy	nAUC
Tran *et al.* [33]	89.40 ± 1.52	89.36 ± 1.25	95.90 ± 0.95
Ours (w/ Gla)	84.37 ± 1.44	85.79 ± 0.42	92.87 ± 1.09
Ours (w/o Gla)	87.69 ± 1.01	89.02 ± 0.89	95.37 ± 0.65

6 Conclusions

In this work, we describe a 3D face detailed reconstruction framework that can run efficiently under occluded scenes. Our method enables unobstructed face image synthesis by concatenating the original face parsing map with the face edge lines map which both are extracted from the input face image in the encoder-decoder network. The experiments on 3D face reconstruction using various datasets have shown that our method can effectively remove eyeglasses with equivalent quality and better accuracy control than the existing methods.

Acknowledgements. This paper is supported by National Natural Science Foundation of China (No. 62072020) and the Leading Talents in Innovation and Entrepreneurship of Qingdao (19-3-2-21-zhc).

References

1. Abrevaya, V.F., Boukhayma, A., Torr, P.H., Boyer, E.: Cross-modal deep face normals with deactivable skip connections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4979–4989 (2020)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: Siggraph, vol. 99, pp. 187–194 (1999)
3. Cheng, S., Tzimiropoulos, G., Shen, J., Pantic, M.: Faster, better and more detailed: 3D face reconstruction with graph convolutional networks. In: Proceedings of the Asian Conference on Computer Vision (2020)
4. Chu, X., Ouyang, W., Li, H., Wang, X.: Structured feature learning for pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4715–4723 (2016)
5. 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)
6. 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). pp. 534–551 (2018)
7. 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)
8. Guo, T., et al.: Residual encoder decoder network and adaptive prior for face parsing. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)

9. Guo, Y., Cai, J., Jiang, B., Zheng, J.: CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. IEEE Trans. Pattern Anal. Mach. Intell. **41**(6), 1294–1307 (2018)
10. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on faces in 'Real-Life' images: detection, alignment, and recognition (2008)
11. Kemelmacher-Shlizerman, I., Basri, R.: 3D face reconstruction from a single image using a single reference face shape. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 394–405 (2010)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV workshops), pp. 2144–2151. IEEE (2011)
14. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_49
15. 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)
16. Lee, G.H., Lee, S.W.: Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6100–6109 (2020)
17. Li, K., et al.: Joint face alignment and 3D face reconstruction with efficient convolution neural networks. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 6973–6979. IEEE (2021)
18. Li, X., Wu, S.: Multi-attribute regression network for face reconstruction. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 7226–7233. IEEE (2021)
19. Liu, P., Han, X., Lyu, M., King, I., Xu, J.: Learning 3D face reconstruction with a pose guidance network. In: Proceedings of the Asian Conference on Computer Vision (2020)
20. Masi, I., Mathai, J., AbdAlmageed, W.: Towards learning structure via consensus for face segmentation and parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5508–5518 (2020)
21. 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)
22. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
23. 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, pp. 2536–2544 (2016)
24. Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 497–500 (2001)

25. Richardson, E., Sela, M., Kimmel, R.: 3D face reconstruction by learning from synthetic data. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 460–469. IEEE (2016)
26. 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
27. 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)
28. Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Scribbler: controlling deep image synthesis with sketch and color. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5400–5409 (2017)
29. 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)
30. Shang, J., et al.: Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. arXiv preprint arXiv:2007.12494 (2020)
31. Te, G., Liu, Y., Hu, W., Shi, H., Mei, T.: Edge-aware graph representation learning and reasoning for face parsing. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 258–274. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_16
32. 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, pp. 2387–2395 (2016)
33. 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)
34. 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, pp. 3935–3944 (2018)
35. Vetter, T., Blanz, V.: Estimating coloured 3D face models from single images: an example based approach. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 499–513. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0054761
36. Wei, Z., Liu, S., Sun, Y., Ling, H.: Accurate facial image parsing at real-time speed. IEEE Trans. Image Process. **28**(9), 4659–4670 (2019)
37. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: a boundary-aware face alignment algorithm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2129–2138 (2018)
38. Ye, L., Zhang, B., Yang, M., Lian, W.: Triple-translation GAN with multi-layer sparse representation for face image synthesis. Neurocomputing **358**, 294–308 (2019)
39. Zhan, F., Zhu, H., Lu, S.: Spatial fusion GAN for image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3653–3662 (2019)
40. Zhang, Y., Zhang, H., Wu, G., Li, J.: Spatio-temporal self-supervision enhanced transformer networks for action recognition. In: IEEE International Conference on Multimedia and Expo (ICME). IEEE (2022)

41. Zhang, Y., Zhang, H., Wu, G., Xu, Y., Shi, Z., Li, J.: TMN: temporal-guided multiattention network for action recognition. In: 2022 26th International Conference on Pattern Recognition (ICPR). IEEE (2022)
42. Zhao, D., Qi, Y.: Generative face parsing map guided 3D face reconstruction under occluded scenes. In: Magnenat-Thalmann, N., et al. (eds.) CGI 2021. LNCS, vol. 13002, pp. 252–263. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-89029-2_20