Credible and Detailed 3D Face Reconstruction in Large Pose

1st Xinyu Li*
School of Information Engineering
Ningxia University
Yinchuan, China
lxy_nxu@163.com

2nd Xitie Zhang*
School of Information Engineering
Ningxia University
Yinchuan, China
saterzhan@163.com

3rd Suping Wu†
School of Information Engineering
Ningxia University
Yinchuan, China
pswuu@nxu.edu.cn

4th Ruijie Peng
School of Information Engineering
Ningxia University
Yinchuan, China
prj105@stu.nxu.edu.cn

5th Kehua Ma
School of Information Engineering
Ningxia University
Yinchuan, China
kehuamm@163.com

6th Xiang Zhang
School of Information Engineering
Ningxia University
Yinchuan, China
zxiang7996@stu.nxu.edu.cn

***Abstract*—The existing monocular methods face huge challenges in reconstructing credible details of non-visible areas in large pose images. Due to the fact that facial details are lost in non-visible areas of large pose images, existing methods lose basis when reconstructing details, resulting in unreliable results. Even if the generative model is used to repair the image first, the reconstruction process is very complex and costly. To this end, we propose an end-to-end and self-supervised RGB to depth method that equates small pose to large pose in UV space to obtain labels for non-visible areas. Then, we infer the depth values of non-visible areas and reconstruct a detailed 3D face. Finally, we render it into a face image and use it alongside the labels for self-supervised training of the network. During inference for large pose image, our method could reconstruct credible details of non-visible areas with a basis, rather than blindly. In addition, coarse and detailed reconstruction have mutually exclusive requirements for training images while coarse reconstruction is often not met, resulting in limited reconstruction accuracy. We propose a mutual exclusion elimination method to solve it, improving its accuracy. Extensive experiments demonstrate that our method could reconstruct precise 3D faces with credible details from large pose images. Our supplementary material is published on: https://github.com/lxy-nxu/ICASSP2025/tree/main**

***Index Terms*—Detailed 3D face reconstruction, Large pose scenarios, Non-visible areas**

I. Introduction

Detailed 3D facial reconstruction from a single image holds a significant position in computer vision, and existing methods [1]–[4] could accurately reconstruct details for images with small poses. They all employ a two-stage reconstruction approach. Initially, they reconstruct a coarse 3D face, then estimate a displacement depth map to represent details, and finally combine both to produce a detailed 3D face. In the coarse reconstruction phase, these methods estimate facial parameters from images and fit them to 3DMM or FLAME models to form a 3D face. 3DMM [5] and FLAME [6] are statistical models learned from a collection of facial scans. There are many coarse 3D face reconstruction methods [7]–[9], [23]–[40] that employ them. In the detailed reconstruction phase, they reconstruct details through the displacement depth map computed by the pix2pix network, decoder, translation network or SFS [10] method. However, for images with large poses, the details in non-visible areas are often missing, leading to incomplete or incorrect reconstructions by existing methods. Due to space limitations, we show it in Fig. 2 of the supplementary materials. The same applies to some subsequent displays. Furthermore, we attempt to apply the generation and diffusion models to repair the missing textures and then conduct the reconstruction, as shown in the section 1.1.2 of the supplementary materials. However, this solution faces more scattered steps and high costs, and the effect is not ideal.

To address this in an end-to-end form, we propose a self-supervised RGB to depth method. We use small pose images as training data and labels, unwrap them into UV maps, and replace the inner half of the face with the outer half before using the network to restore the original content of the inner half. We then reconstruct a detailed 3D face from it, render it into images, and use these images for self-supervised training of the network. We found that recovering details directly in the RGB domain is particularly challenging because RGB contains too much information. Consequently, we shifted the recovery process from the RGB domain to the depth domain, where there is less redundant information, allowing the network to recover details more accurately. During inference, we unwrap the large pose input images into UV maps and replace the non-visible parts with visible ones, enabling the network to directly infer the original depth values of the non-visible areas and thus reconstruct credible details.

Existing detailed 3D face reconstruction methods usually utilize datasets such as Celeb A [11], VGGFace2 [12], BUPT-Balancedface [13], and VoxCeleb2 [14] for training. We

* represents equal contributions and † represents corresponding author. This work is supported by Ningxia Natural Science Foundation Project under Grant 2024AAC02012 and in part by the National Natural Science Foundation of China under Grant(62062056).

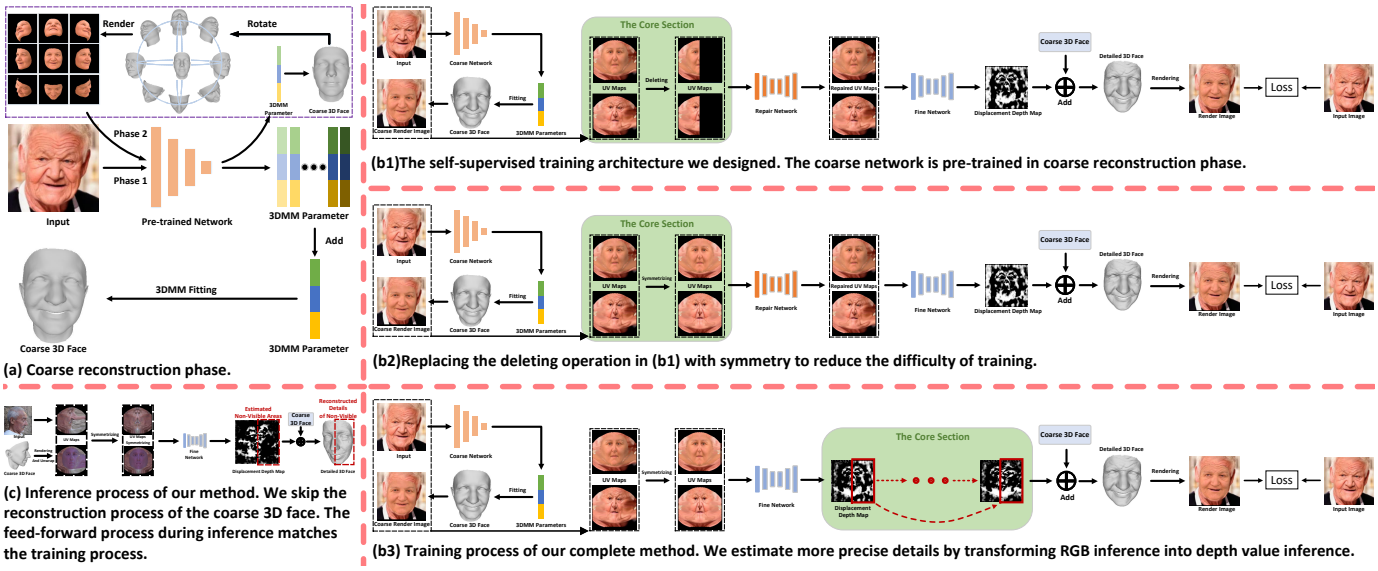

Fig. 1. (a) and (b) are the coarse and detailed reconstruction phases, respectively. (b1), (b2), and (b3) show the progressive design step, while (b3) is the form of the final application. (c) is the inference process.

TABLE I
WE ANALYZE THE PROPORTION OF POSES WITHIN THE FOUR DATASETS.

Dataset	Test Number	Small and Medium Pose			Large Pose		
		$[0°, 15°]$	$(15°, 30°]$	$(30°, 45°]$	$(45°, 60°]$	$(60°, 75°]$	$(75°, 90°]$
Four Datasets	4,489,391	2608103	1139634	456953	192535	78318	13848
		93.658%			6.342%		

analyzed the number and proportion of large pose faces within these datasets based on their yaw angles, with the results displayed in Table 1. It is evident that large poses constitute only a small fraction (6.342%) of the datasets. To investigate the impact of the proportion of large poses on reconstruction, we conducted further verification in section 2.2 of the supplementary materials. It suggests that the 6.342% proportion of large poses in existing methods is highly detrimental to coarse network learning. As learning is significantly restricted, it becomes challenging to improve the accuracy of reconstructions. This highlights that the demands for coarse and detailed reconstructions are mutually exclusive: detailed reconstruction necessitates training images with small poses to learn complete details, whereas coarse reconstruction requires training images with large poses to comprehend the rich 3D structure of faces. To solve the mutually exclusive problem, we render additional large pose images by controlling the pose parameters and using them to retrain the network, thereby enhancing the reconstruction accuracy of existing methods.

II. METHOD

In the coarse reconstruction phase, our goal is to eliminate the mutually exclusive problem. We leverage the reconstruction network from existing methods to generate images of large poses and use these images to retrain the network, effectively increasing the proportion of large poses in the training dataset. As shown in Fig. 1(a), after estimating the 3DMM parameters, we fit them to a 3D face and then apply multiple large yaw poses to generate several large pose 3D faces. These are subsequently rendered into multiple images. Since each batch size includes multiple input images, to identify which rendered images correspond to which face, we add an original image at the beginning of each set of rendered images as an identity guide before inputting them together into the network for retraining. The network estimates the 3DMM parameters for each image, and we sum and merge the parameters belonging to the same input image, using them to fit the 3D face of that input image. As the proportion of large-pose training images in the reconstruction network increases, following the trend in supplementary materials Fig. 8, existing methods' reconstruction error will further decrease.

In the detailed reconstruction phase, we propose an end-to-end and self-supervised RGB to depth method. Firstly, to infer credible details for non-visible areas, it is essential to train the inference network (UNet) with real labels. However, training datasets with multiple poses are scarce, meaning that high-quality labels are difficult to obtain. To this end, we design a self-supervised training architecture as illustrated in Fig. 1(b1), where a single image serves both as training data and the real label. Our approach involves deleting the inner half of the face from the UV unwrap of a small pose input image and then employing UNet as a repair network to infer the facial details of the deleted area. Subsequently, we render the reconstructed details back into the image and use the input image to supervise the rendered image, thus providing the fine network (UNet) with real labels for training. The reason for using the inner half of the face from the UV unwrap in our approach is that when inferring details from images with large poses, this region corresponds to the non-visible areas that need to be inferred. We use images with small poses as input because faces with minor pose variations can provide reliable labels for inference, and existing training

TABLE II
QUANTITATIVE RESULTS OF INTEGRATING OUR METHOD INTO SOTA METHODS.

Methods / 3D NME	MICC Small Pose	MICC Large Pose	NOW Small Pose	NOW Large Pose	3dr Large Pose	3ds Large Pose
Chen20 [3]	2.510±2.176	3.177±2.618	2.824±2.468	3.559±2.928	2.413±1.915	2.459±1.778
Chen20 + Our	2.352±2.010	2.476±2.109	2.317±2.019	2.341±1.992	2.225±1.974	2.178±1.779
DECA [2]	2.347±2.244	2.363±2.262	2.395±2.433	2.602±2.392	2.210±2.015	2.141±1.695
DECA + Our	2.309±2.095	2.344±2.139	2.368±2.211	2.514±2.215	2.176±2.126	2.191±1.887

datasets are primarily composed of such images (verified in section 2.2.3 of the supplementary materials). When the input image is of a large pose, it is not a concern, since the details of the non-visible areas are absent in both the input and the rendered images, the loss is zero, which does not affect the training process. Based on the prior knowledge that faces are fundamentally symmetrical, we have optimized the deletion process by replacing the inner half of the face with the outer half. This allows the network to infer the correct details based on symmetry, thereby simplifying the training process. The workflow is illustrated in Fig. 1(b2). However, inferring facial details in the RGB domain is challenging due to the multitude of information contained beyond just facial details, which could severely disrupt the inference process. To address this, we propose a transition from RGB to depth. We posit that depth maps primarily contain facial details and, being single-channel, are inherently simpler to estimate than RGB values. Consequently, we transform the task of RGB inference into depth value inference, significantly reducing interference and enabling more accurate detail inference. The training pipeline of our complete methodology is illustrated in Fig. 1(b3), demonstrates a significant reduction in complexity compared to RGB inference, which requires two networks. Our approach needs only one. As shown in Fig. 1(c), during inference, we also apply symmetry to the non-visible areas of the UV map. The network then infers the correct details based on this symmetry, resulting in a 3D face with complete and accurate details. The total loss function is shown in the section 1.2 of the supplementary materials.

III. EXPERIMENTS

We performed quantitative evaluation using the MICC Florence [15], NOW validation [16], and 3dMDLab [17] (including 3dr and 3ds subsets) datasets. The details are written in the section 2.3.2 of the supplementary materials. To validate the effectiveness of our proposed method, we integrated it into SOTA models. The most recent SOTA for detailed 3D face reconstruction is HRN [18], but since its training code is not publicly available, we selected the latest SOTA models with available training codes, "DECA" [2] and "Chen20" [3], which are also highly influential. We write the complete implementation details of the coarse and detailed reconstruction phase in sections 2.3.2 and 1.3 of the supplementary materials.

A. Quantitative Experiment

From the results shown in Table 2, it is evident that the proportion of large pose training data in Chen20 was significantly increased after integrating our method, which correspondingly led to a substantial reduction in 3D normalized mean error (NME) for reconstruction. This confirms our hypothesis that the small proportion of large poses in existing methods is detrimental to network learning, and that increasing the proportion of large poses can enhance reconstruction accuracy. Additionally, incorporating our method into DECA also resulted in a reduction in overall 3D NME. Both outcomes collectively indicate that the reconstruction networks of existing methods are indeed constrained by the training data, improving the training dataset enhanced the network's reconstruction precision. This validates our approach and proves the effectiveness of our method.

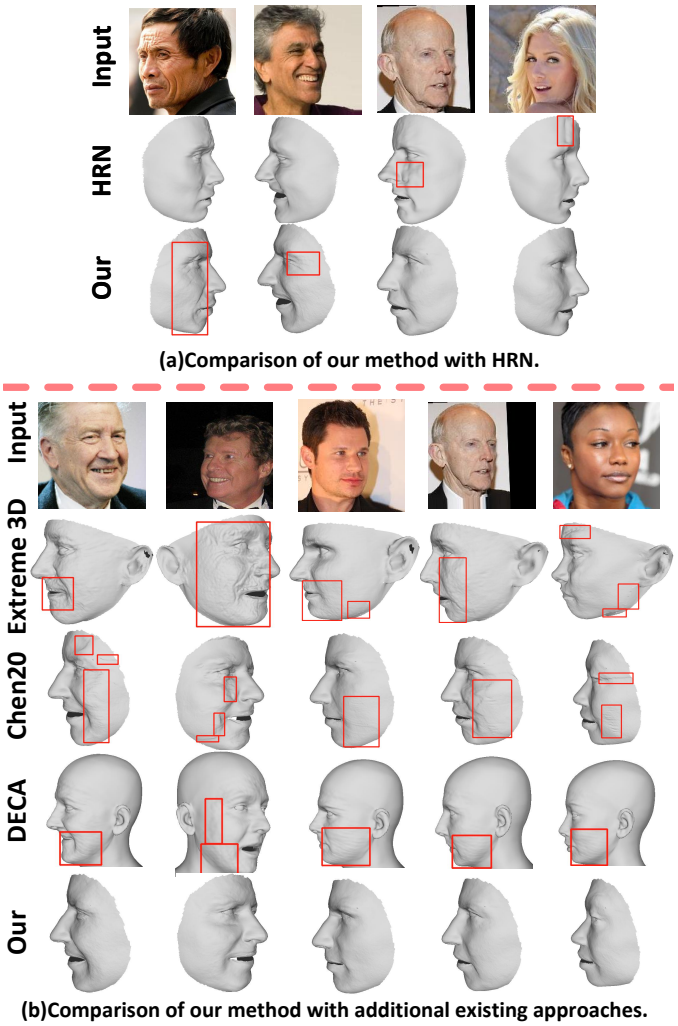

Fig. 2. Comparison of our method with the existing detailed 3D face reconstruction approaches.

B. Qualitative Experiment

Existing methods yield two types of outcomes: detail loss and detail disarray. In the first two columns of Fig. 2(a), we demonstrate the comparative results in cases of detail loss. It is evident that HRN does not reconstruct invisible details, which, although it prevents errors, also results in details loss in the 3D face. In contrast, our method can reconstruct details in non-visible areas, such as wrinkles. Additionally, the latter two columns of Fig. 2(a) show that HRN exhibits disarray when estimating the edges of large pose faces, while our method could reconstruct more credible details. More examples of the second type of outcome are displayed in Fig. 2(b), where existing methods exhibit detail disarray, whereas our results do not.

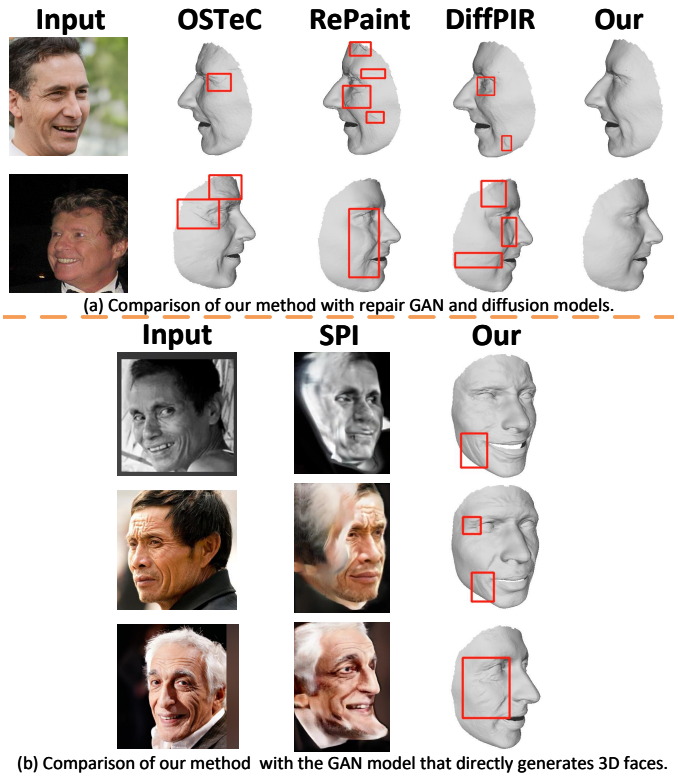

Fig. 3. Comparison of our method with GAN and diffusion models.

We also conducted a comparison of the inference approaches of GAN [19], [20] and diffusion [21], [22] models, which is presented in Fig. 3. For the three methods related to image repair, because they are prone to infer incorrect content, the details they reconstruct are often disordered, as shown in the red boxes in Fig. 3(a). In contrast, our method reconstructs more credible details. As for the method that directly generate 3D face, it often suffers from distortions in the basic shape, significantly reducing the accuracy of the reconstruction, as demonstrated in Fig. 3(b). Conversely, our results not only maintain accurate shapes but also further reconstruct the wrinkle details in the non-visible areas, as highlighted in the red boxes in Fig. 3(b). These results collectively highlight the deficiencies of existing methods in reconstructing large pose faces and also validate our self-supervised and RGB to depth approach's ability to train the network to infer details in non-visible areas. Consequently, our method offers an end-to-end solution to the problem. In addition, since they are not end-to-end architecture solutions, applying them requires an extra repair step before performing normal reconstruction using existing detailed reconstruction methods. This leads to additional overhead. The average GPU memory utilization for these four models is 33,160 MB, and the average inference time is 2,196 seconds per image.

C. Ablation Study

To validate the end-to-end and self-supervised RGB to depth method for credible and detailed reconstruction independently, we conducted an ablation study as shown in Fig. 4. 'Our w/o' denotes the removal of our approach. As shown by the red circles in the displacement depth maps, it results in disordered inferred details, which are further corroborated by the red boxes on detailed 3D faces. This demonstrates that our method achieves our design objectives, specifically inferring credible details for non-visible areas to prevent detail loss or disarray.

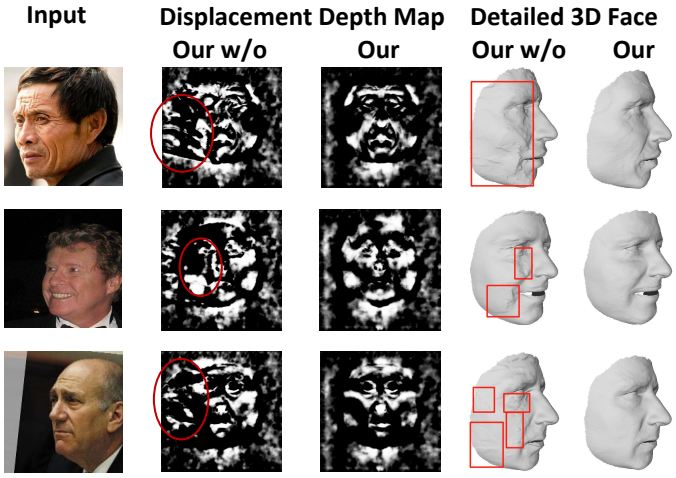

Fig. 4. Ablation study results. 'Our w/o' denotes the removal of our approach.

IV. Conclusion

Since the details in the non-visible areas of large pose images are missing, existing methods could not reconstruct them credibly. We propose a self-supervised RGB to depth method to infer the details of these non-visible areas, achieving credible detail reconstruction. The accuracy of coarse reconstruction is severely restricted due to the influence of mutual exclusion problems. Therefore, we propose a mutual exclusion elimination method. It renders more large pose images by controlling pose parameters and retrains the network, thereby enhancing the precision of reconstructions. By combining these two approaches, we could reconstruct precise 3D faces with credible details from large pose images.

REFERENCES

[1] Lei B, Ren J, Feng M, et al. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 394-403.

[2] Feng Y, Feng H, Black M J, et al. Learning an animatable detailed 3D face model from in-the-wild images[J]. ACM Transactions on Graphics (ToG), 2021, 40(4): 1-13.

[3] Chen Y, Wu F, Wang Z, et al. Self-supervised learning of detailed 3d face reconstruction[J]. IEEE Transactions on Image Processing, 2020, 29: 8696-8705.

[4] Tran A T, Hassner T, Masi I, et al. Extreme 3D Face Reconstruction: Seeing Through Occlusions[C]//CVPR. 2018, 1: 2.

[5] Blanz V, Vetter T. A morphable model for the synthesis of 3D faces[M]//Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023: 157-164.

[6] Li T, Bolkart T, Black M J, et al. Learning a model of facial shape and expression from 4D scans[J]. ACM Trans. Graph., 2017, 36(6): 194:1-194:17.

[7] Deng Y, Yang J, Xu S, et al. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019: 0-0.

[8] Tewari A, Zollhofer M, Kim H, et al. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction[C]//Proceedings of the IEEE international conference on computer vision workshops. 2017: 1274-1283.

[9] Jiang L, Wu X J, Kittler J. Dual attention mobdensenet (damdnet) for robust 3d face alignment[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019: 0-0.

[10] Zhang R, Tsai P S, Cryer J E, et al. Shape-from-shading: a survey[J]. IEEE transactions on pattern analysis and machine intelligence, 1999, 21(8): 690-706.

[11] Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3730-3738.

[12] Cao Q, Shen L, Xie W, et al. Vggface2: A dataset for recognising faces across pose and age[C]//2018 13th IEEE international conference on automatic face gesture recognition (FG 2018). IEEE, 2018: 67-74.

[13] Wang M, Deng W, Hu J, et al. Racial faces in the wild: Reducing racial bias by information maximization adaptation network[C]//Proceedings of the ieee/cvf international conference on computer vision. 2019: 692-702.

[14] Chung J S, Nagrani A, Zisserman A. Voxceleb2: Deep speaker recognition[J]. arXiv preprint arXiv:1806.05622, 2018.

[15] Bagdanov A D, Del Bimbo A, Masi I. The florence 2d/3d hybrid face dataset[C]//Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding. 2011: 79-80.

[16] Sanyal S, Bolkart T, Feng H, et al. Learning to regress 3D face shape and expression from an image without 3D supervision[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7763-7772.

[17] Booth J, Roussos A, Ververas E, et al. 3D reconstruction of "in-the-wild" faces in images and videos[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 40(11): 2638-2652.

[18] Lei B, Ren J, Feng M, et al. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 394-403.

[19] Gecer B, Deng J, Zafeiriou S. Ostec: One-shot texture completion[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7628-7638.

[20] Yin F, Zhang Y, Wang X, et al. 3d gan inversion with facial symmetry prior[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 342-351.

[21] Lugmayr A, Danelljan M, Romero A, et al. Inpainting using denoising diffusion probabilistic models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11461-11471.

[22] Zhu Y, Zhang K, Liang J, et al. Denoising diffusion models for plug-and-play image restoration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 1219-1229.

[23] Zhu X, Lei Z, Liu X, et al. Face alignment across large poses: A 3d solution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 146-155.

[24] Feng Y, Wu F, Shao X, et al. Joint 3d face reconstruction and dense alignment with position map regression network[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 534-551.

[25] Zhou Z, Li L, Wu S, et al. Replay attention and data augmentation network for 3-D face and object reconstruction[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2023, 5(3): 308-320.

[26] Ren X, Lattas A, Gecer B, et al. Facial geometric detail recovery via implicit representation[C]//2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, 2023: 1-8.

[27] Yu R, Saito S, Li H, et al. Learning dense facial correspondences in unconstrained images[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 4723-4732.

[28] Li X, Wu S. Multi-attribute regression network for face reconstruction[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 7226-7233.

[29] Li L, Li X, Wu K, et al. Multi-granularity feature interaction and relation reasoning for 3d dense alignment and face reconstruction[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 4265-4269.

[30] Wang X, Li X, Wu S. Graph structure reasoning network for face alignment and reconstruction[C]//MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part I 27. Springer International Publishing, 2021: 493-505.

[31] Wang J, Zhang S, Song F, et al. Exploring occlusion-sensitive deep network for single-view 3D face reconstruction[C]//2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022: 1821-1825.

[32] Li L, Liu F, Wang J, et al. Exploiting global and instance-level perceived feature relationship matrices for 3D face reconstruction and dense alignment[J]. Engineering Applications of Artificial Intelligence, 2024, 131: 107862.

[33] Zhu C, Wu S, Yu Z, et al. Multi-agent deep collaboration learning for face alignment under different perspectives[C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 1785-1789.

[34] Zhu C, Li X, Li J, et al. Spatial-temporal knowledge integration: Robust self-supervised facial landmark tracking[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 4135-4143.

[35] Liu F, Zeng D, Li J, et al. Cascaded regressor based 3D face reconstruction from a single arbitrary view image[J]. arXiv preprint arXiv:1509.06161, 2015, 2.

[36] Liu F, Zeng D, Zhao Q, et al. Joint face alignment and 3d face reconstruction[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer International Publishing, 2016: 545-560.

[37] Yu X, Huang J, Zhang S, et al. Face landmark fitting via optimized part mixtures and cascaded deformable model[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(11): 2212-2226.

[38] Burgos-Artizzu X P, Perona P, Dollár P. Robust face landmark estimation under occlusion[C]//Proceedings of the IEEE international conference on computer vision. 2013: 1513-1520.

[39] Cao X, Wei Y, Wen F, et al. Face alignment by explicit shape regression[J]. International journal of computer vision, 2014, 107: 177-190.

[40] Li L, Liu F, Wang J, et al. Exploiting global and instance-level perceived feature relationship matrices for 3D face reconstruction and dense alignment[J]. Engineering Applications of Artificial Intelligence, 2024, 131: 107862.