SFPFR: Self-supervised Facial Paralysis Face Reconstruction Under Few Views

Yaru Qiu[1,*], Xinru Wang[2,*], Jianfang Zhang[1], Bo Liu[3] , Peng Bai[4,✉], Yuanyuan Sun[1,✉]

[1] School of Data Science, Qingdao University of Science and Technology
[2] Beijing University of Chinese Medicine
[3] Beihang University
[4] School of Acupuncture-moxibustion and Tuina, Beijing University of Chinese Medicine
yysun@qust.edu.cn, baipeng202305@163.com

Abstract. 3D face reconstruction methods exhibit significant limitations when applied to pathological cases such as facial paralysis, due to inherent challenges including asymmetric motion and non-linear muscle dynamics. To address these gaps, we propose SFPFR, a self-supervised framework for facial paralysis 3D face reconstruction leveraging 1-3 viewpoints. We first propose a self-supervised learning paradigm integrating reconstruction loss, multi-view consistency loss, and a Mamba-based temporal loss to reconstruct 3D face without ground-truth; then, a partitioned dynamic fusion module that adaptively weights multi-view features ensuring precise geometric reconstruction and pathological detail preservation; last, we propose FPD-100, the first multi-view video dataset for facial paralysis, comprising 30,000 frames from 100 patients of 3 views. Extensive experiments validate SFPFR's superiority, achieving state-of-the-art PSNR (27.74) and FID (37.13). It enables clinical applications in severity assessment, rehabilitation monitoring, and treatment planning, while the dataset and code will be open-sourced to catalyze research in pathological facial analysis.

Keywords: 3D Face Reconstruction · Facial Paralysis · Self-supervision

1 Introduction

Facial paralysis is the loss or weakening of facial muscle function, causing asymmetric expressions. Assessment is crucial for diagnosis, treatment, and rehabilitation, helping physicians determine the type and severity of facial paralysis. Currently, assessments rely on subjective judgment using rating scales like the House-Brackmann scale, which is time-consuming and labor-intensive. 3D face reconstruction, which generates 3D facial models from 2D images, shows great promise in medical diagnostics and virtual reality. However, using professional 3D scanning equipment for facial paralysis patients faces challenges, including

Yaru Qiu* and Xinru Wang* contributed equally to this work.

high costs, lengthy scanning times, and the need for patient cooperation. Recently, 3D face reconstruction methods based on monocular 2D images have been explored in clinical assessments of facial paralysis [1], with notable progress in reconstructing identity features and basic facial geometry.

However, facial reconstruction specifically for facial paralysis still faces three core challenges. Challenge 1: Pathological Facial Asymmetry and Motion Distortion. Most mainstream 3D face reconstruction methods [2,3] use traditional 3D deformable face models (3DMM [4]) and focus on estimating parameters from models like BFM [5], FaceWarehouse [6], or FLAME [7]. These methods rely on linear combinations of facial shapes (e.g., BlendShapes [8]) to represent expression changes, which fails to capture the nonlinear muscle dynamics in facial paralysis patients. As a result, they struggle to reconstruct facial asymmetry and abnormal muscle movements. While approaches like Smirk [9] attempt to address asymmetry using balanced loss functions and neural rendering, they struggle to capture the subtle details of facial paralysis, resulting in poor reconstruction quality and an incomplete representation of the affected facial movements. Our self-supervised framework integrates the proposed losses and temporal features for accurate asymmetric reconstruction. Challenge 2: Lack of Facial Paralysis Motion Images and 3D Ground Truth Data. Privacy concerns and data collection difficulties lead to a severe shortage of authentic 2D and 3D facial paralysis data, hindering the training of existing reconstruction methods. To address this critical gap, we introduce the FPD-100 dataset, providing essential pathological motion data. Although self-supervised frameworks using perceptual or differentiable rendering loss [9–12] have been developed, they are trained on healthy population data and struggle to generalize to facial paralysis, particularly in handling asymmetry and pathological details. This underscores the need for models tailored for facial paralysis reconstruction. Challenge 3: Insufficient View Compatibility. Most algorithms are restricted by fixed modes, making it difficult to handle varying input images (single or multi-view). Single-view methods fail to capture the multidimensional indicators needed for facial paralysis assessment, while multi-view algorithms [13] struggle with single-view inputs. Our partitioned dynamic fusion module handles any number of input views adaptively.

To address these three challenges, we propose SFPFR, a framework achieving high-fidelity facial reconstruction of facial paralysis patients from a few viewpoint images. Our contributions are summarized as follows:

- A self-supervised framework that integrates reconstruction loss, multi-view consistency loss, and temporal action feature capture module to reconstruct asymmetric facial paralysis models without ground truth data.
- A large-scale multi-view video dataset (FPD-100) containing 30,000 frames from 100 facial paralysis patients performing standardized facial movements, captured simultaneously from three calibrated views.
- A partitioned dynamic fusion module that adaptively weights multi-view features, enabling precise geometric reconstruction and pathological detail preservation.

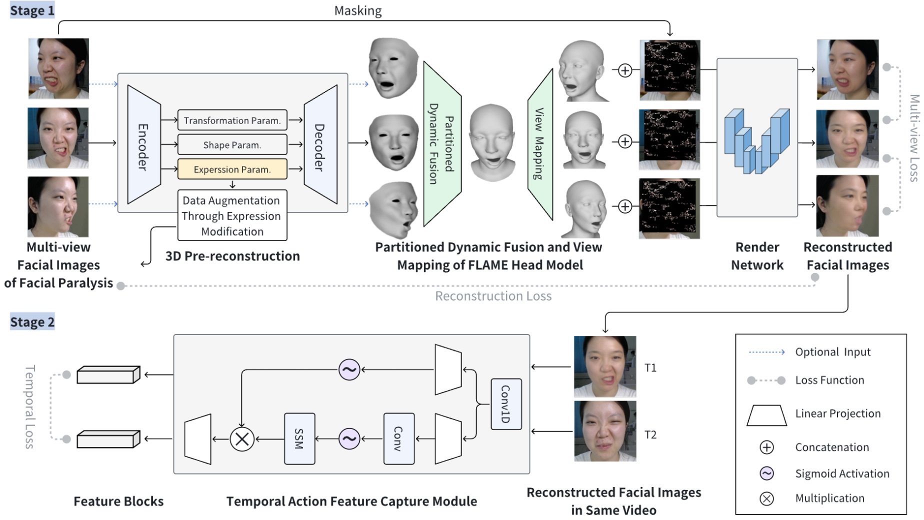

Fig. 1. Our proposed SFPFR Architecture. Stage 1 reconstructs and fuses views of the same person from different perspectives. Stage 2 conducts temporal optimization using multiple video frames of the same individual. This model can handle both single-view and multi-view inputs. For single-view input, the Dynamic Fusion module is inactive.

2 Method

As shown in Figure 1, our method processes single-view and multi-view facial images of facial paralysis using a two-stage training system. The first stage (Section 2.1) involves viewpoint reconstruction and image rendering, while the second stage (Section 2.2) employs the temporal action feature capture module to further capture the unique temporal facial movement characteristics of facial paralysis patients.

2.1 Stage 1 of SFPFR

3D Pre-reconstruction Module As shown in Figure 1, the facial image (single-view I_{sg} or multi-view I_i, $i \in \{0,1,2\}$) is processed through the 3D pre-reconstruction module to generate a base facial model consisting of 5023 vertices (S_{sg} or $S_i(i \in \{0,1,2\})$). The process begins by applying the FLAME algorithm [7] to extract facial parameters (focusing on key parameters related to eyelids, lower face and jaw rotation). Subsequently, an Encoder (E, comprising three MobileNetV3 networks [14]) is employed to predict expression parameters φ, shape parameters α, and global transformation parameters θ through regression (as shown in Equation 1). Finally, according to Equation 2, a Decoder (R, constructed with multiple deconvolution layers) utilizes these parameters $(\theta, \alpha, \varphi)$ to generate the basic 3D facial model S.

$$\theta = E_\theta\left(I_i\right), \alpha = E_\alpha\left(I_i\right), \varphi = E_\varphi\left(I_i\right) \tag{1}$$

$$S = R(\alpha, \beta, \varphi) \tag{2}$$

Partitioned Dynamic Fusion Module and View Mapping Module We propose a viewpoint processing strategy to accommodate varying numbers of input facial images, enabling three-viewpoint rendering (left, center, right) from both single-view and multi-view inputs.

(1) Partitioned Dynamic Fusion Module(PDFM) The 3D pre-reconstruction module generates three viewpoint facial models $S_i (i \in \{0, 1, 2\},\ S_i \in \mathbb{R}^{5023})$. Frontal viewpoints excel at reconstructing eye areas but cannot effectively capture cheek asymmetry, which is better expressed by lateral models. We developed the PDFM to address this issue. After aligning the three viewpoint models, we implemented a dynamic weight redistribution factor $w(x)$ that enables weighted fusion from different directions.

$$w(x) = \begin{cases} 1 - \frac{x - x_{\text{min}}}{x_{\text{mid}} - x_{\text{min}}}, x \in [x_{\text{min}}, x_{\text{mid}}) \\ \frac{x - x_{\text{mid}}}{x_{\text{max}} - x_{\text{mid}}}, x \in [x_{\text{mid}}\ , x_{\text{max}}] \end{cases} \tag{3}$$

In the facial coordinate system, x_{min} represents the leftmost point on the x-axis, x_{mid} denotes the central position of the x-axis, and x_{max} indicates the rightmost point.The fusion model S at the point $S(x, y, z)$ is calculated as follows:

$$S(x, y, z) = \begin{cases} w(x) S_0(x, y, z) + (1 - w(x)) S_1(x, y, z), x \in [x_{\text{min}}, x_{\text{mid}}) \\ w(x) S_2(x, y, z) + (1 - w(x)) S_1(x, y, z), x \in [x_{\text{mid}}, x_{\text{max}}] \end{cases} \tag{4}$$

where $w(x)$is the weight function, S_1 represents the front view, S_0 represents the left view, and S_2 represents the right view. This paper introduces dynamic weights in the PDFM, using a spatially varying weight function to adaptively fuse three view models, determining the priority of the asymmetric side view of the facial paralysis cheek while maintaining the frontal accuracy of the eye region.

(2) View Mapping Module (VMM) Single-view facial reconstruction often suffers from geometric distortions, particularly around the mouth area, leading to inaccuracies in the model. To address this, our VMM applies multi-view consistency constraints to single-view inputs. By using rotation matrices, it generates frontal, left, and right views of the 3D model, which are then processed through a rendering module. This approach ensures consistent and accurate shape reconstruction during optimization, even when only a single-view input is available, enhancing the model's overall quality. Thus, the multi-view consistency loss also enforces geometric alignment across different rendered views to ensure the accurate reconstruction of asymmetric features, such as the mouth deviation in facial paralysis.

Render Module Unlike traditional self-supervision paradigms, we adopt an analysis-by-neural-synthesis-based self-supervision mechanism [9] to synthesize reconstruction facial images from input viewpoints. This module adopts a U-Net architecture design, with the goal of performing three-dimensional rendering reconstruction on the original input image I_i. We randomly sample facial regions, retaining only about 5% of the pixels as key features for training to guide reconstruction, encouraging the network to rely more on the model S for rendering and reconstruction, thereby enhancing the accuracy of the geometric structure.

$$I_i{}' = U\left(S_i \oplus M\left(I_i\right)\right) \tag{5}$$

Where $\oplus$ stands for connection, M represents the masking operation and I_i' denotes the reconstructed facial image, which will be used in the reconstruction loss to optimize the model's geometric shape. Additionally, the parameter-separation optimization strategy updates only the parameters of the expression encoder E_ψ in Equation 1, by applying augmentation through expression modification(as referenced in [9]). The action transformation loss further ensures the stability and effectiveness of the reconstruction process, enhancing the generalization of facial action reconstruction. Notably, the action transformation loss is the same as the multi-view loss applied to the image after expression change, as shown in Equation 6.

Multi-view loss We introduced a multi-view consistency loss $L_{\text{multi-view}}$ (Equation 6).This constraint enhances reconstruction precision from different views, effectively capturing facial asymmetry and subtle muscle movements in facial paralysis patients, and resolving the issue of asymmetry when reconstructing the faces of patients with facial paralysis.

$$L_{\text{multi-view}} = \frac{1}{M}\sum_{j\neq k} ||\widehat{I_j} - \widehat{I_k}||_2^2 \tag{6}$$

Where $\widehat{I_j}$ refers to the rendered image from the j-th viewpoint, and M denotes the number of rendered image pairs.

Reconstruction Loss The VGG loss [15] serves a similar purpose as the photometric loss but accelerates convergence during the early stages of training. To suppress unrealistic expressions and ensure more natural results, we also applied L_2 regularization to the expression parameters ψ. Consequently, the final image reconstruction loss combines perceptual loss and expression regularization, as follows:

$$L_{\text{reconstruction}} = L_{\text{vgg}} + \lambda L_{\text{reg}} = \sum_{l} \left|\Gamma_l\left(I'\right) - \Gamma_l\left(I\right)\right|_1 + \lambda|\psi|_2^2 \tag{7}$$

where $\Gamma(\cdot)$ represents the VGG perceptual encoder. λ is the hyperparameter used to balance the perceptual loss and regularization terms, and in this experiment, it is set to 0.35.

2.2 Stage 2 of SFPFR: Temporal Action Feature Capture Module and Temporal loss

To capture specific action features (e.g., grimaces) and improve the quality of facial paralysis movements in reconstructed images, we designed a Mamba-based [16] temporal action feature capture module that retains non-smooth motion transitions (e.g., muscle spasms) while ensuring consistency in smoother movements, simulating the irregular muscle behavior of facial nerve paralysis.

We also introduced a temporal loss to address feature drift during training. This mechanism maps the reconstruction facial image sequence $X = \{x_1, x_2, \ldots, x_t\}$ through conv1D projection to a high-dimensional feature space, forming a feature sequence F_t that is subsequently input into a state space model for temporal modeling. We employ temporal loss to constrain the smoothness and consistency of the sequence along the temporal dimension, ensuring natural and coherent expression changes. Additionally, we implement dynamic weighting to accommodate sudden movements and non-smooth actions characteristic of facial paralysis patients.

$$L_{\text{temporal}} = \frac{1}{n-1} \sum_{t=1}^{n-1} \left\| \operatorname{Mamba}\left(F_t\right) - \operatorname{Mamba}\left(F_{t+1}\right) \right\|_2^2 * w_t \tag{8}$$

Where $w_t = \exp(-\alpha \left\| \operatorname{Mamba}\left(F_t\right) - \operatorname{Mamba}\left(F_{t+1}\right) \right\|_2$, and the weight is adjusted based on the feature difference between adjacent frames. For changes with larger magnitude, w_t becomes smaller, allowing for non-smooth transitions. Here, α is a hyperparameter controlling the sensitivity of weight changes.

In summary, the total loss used in SFPFR is composed of the formulas (6, 7, 8). Where ν,υ,ρ represent loss weight parameters.

$$L_{\text{total}} = \nu L_{\text{multi-view}} + \upsilon L_{\text{renconstruction}} + \rho L_{\text{temporal}} \tag{9}$$

3 Experiments and Results

3.1 Datasets and Assessment Indicators

We evaluated our method using the FPD-100 dataset, which we constructed from 30,000 facial images of 100 facial paralysis patients captured from three viewpoints. From this dataset, 2,000 images from 10 independent subjects—distinct from those in the training set and covering diverse severity levels of facial paralysis with one frame extracted per second—were reserved as the test set. Additionally, we employed the Facescape [17] dataset for further validation. Performance was assessed using PSNR, SSIM [18], and FID [19] metrics, as well as vertex error to evaluate 3D reconstruction accuracy. This study was ethically approved under reference number 2024DZMEC-466-01.

3.2 Implementation Details

We implemented our method in the PyTorch framework using 4 NVIDIA RTX A10 GPUs. We used a batch size of 12, with loss weights $\nu = 0.4, \upsilon = 0.4, \rho = 0.2$ applied during training. In the core phase, we trained SFPFR (Ours) for 250,000 iterations with a learning rate of 1e-3 and cosine annealing with restart at each epoch.

3.3 Results

Quantitative Results Quantitative comparisons with state-of-the-art methods (Smirk [9], Deep3DFace [20] and 3DFFA-v3 [21], and MVF-Net [13]) confirm SFPFR's superior performance. As shown in Table 1, on the FPD-100 dataset, our method achieved the highest PSNR (27.74) and FID (37.13) with competitive SSIM (0.89), demonstrating unique advantages in reconstructing pathological features. On the FaceScape dataset, our approach significantly outperformed baseline methods with 46.98% of vertices having errors <2mm for single-view reconstruction and 48.54% of vertices having errors <2mm for multi-view reconstruction, surpassing specialized multi-view method, MVF-Net. This consistent excellence across both datasets and in both single-view and multi-view scenarios demonstrates our method's robustness, scalability, and practical value for high-precision reconstruction applications.

Table 1. Quantitative comparison of reconstruction quality on FPD-100 and FaceScape datasets under different view settings.

Datasets	Setting	Methods	PSNR ↑	SSIM ↑	FID ↓
FPD-100	single-view	3DFFA-v3	25.73	**0.89**	37.89
		Deep3Dface	24.46	0.83	45.15
		Smirk	26.52	0.84	38.41
		Ours	**27.74**	0.88	**37.13**
			<0.5mm(%) ↑	<2mm(%) ↑	Avg.(mm) ↓
FaceScape	single-view	Smirk	5.12	46.34	3.14
		3DFFA-v3	5.29	45.62	**2.98**
		Ours	5.24	46.98	3.03
FaceScape	multi-view	MVF-Net	4.34	31.42	7.80
		Ours	**5.64**	**48.54**	2.99

Qualitative Results The results in Figure 2 and 3 highlight our method's advantages. SFPFR accurately reconstructs oral deformations in facial paralysis patients, preserving critical details like drooping mouth corners and mouth deviation, while other methods exhibit shape distortion and detail blurring. Our approach maintains high realism across different age groups and facial features, with particularly accurate facial contour reconstruction.

Table 2. Ablation study results. A: Multi-view consistency loss. B: Temporal action feature capture module. C: Expression modification. D: Partitioned dynamic fusion.

No.	Module A	Module B	Module C	Module D	PSNR↑	SSIM↑	FID↓
1		✓	✓		27.74	0.88	37.13
2	✓	✓	✓		27.93	**0.89**	35.89
3	✓		✓		26.54	0.81	36.93
4	✓	✓			27.53	0.83	37.06
5	✓	✓	✓	✓	**28.20**	**0.89**	**35.35**

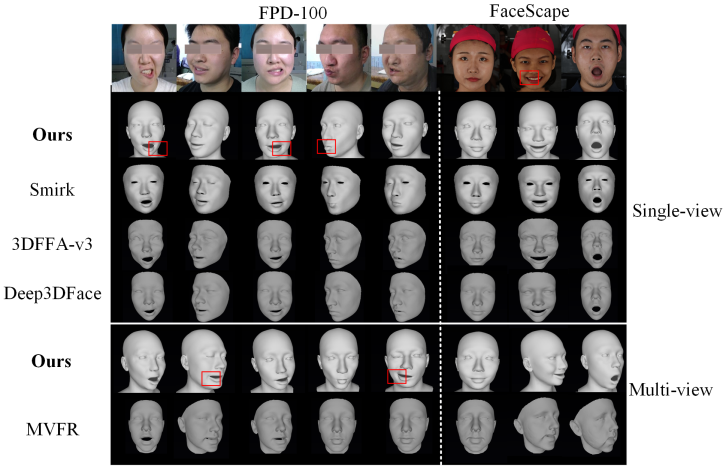

Fig. 2. Comparison of mesh reconstruction quality between SFPFR and other methods.

Ablation Study We conducted ablation studies to validate the effectiveness of each module in the reconstruction process, as shown in Table 2. The results from experiments 1, 2, and the comparison between experiments 2 and 5 demonstrate that Modules A and D improved PSNR and SSIM, enhancing facial movement detail capture. Additionally, the comparison between experiments 2 and 4 shows that Module C significantly improved reconstruction quality, while experiments 2 and 3 confirm that Module B enhanced temporal coherence and captured facial expression changes in facial paralysis patients.

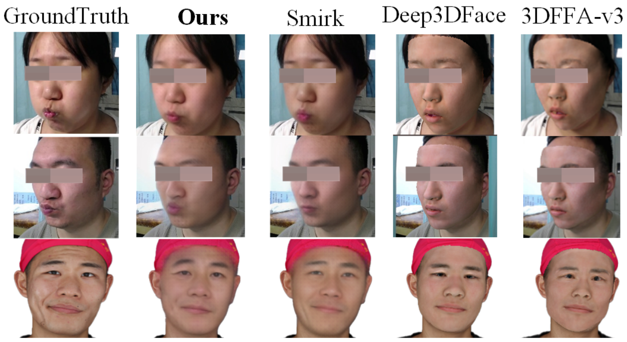

Fig. 3. Visual comparison of rendered facial reconstructions between SFPFR and other methods.

4 Conclusion

We presents SFPFR, a novel self-supervised framework for 3D facial reconstruction of facial paralysis patients using 1-3 viewpoints. Our framework addresses three critical challenges through key innovations: (1) a self-supervised learning paradigm with multi-view consistency loss and Mamba-based temporal loss for capturing asymmetric pathological features without ground truth; (2) a partitioned dynamic fusion module for precise geometric reconstruction; and (3) a temporal action feature capture module for ensuring temporal coherence. Extensive experiments on our FPD-100 dataset and public benchmarks demonstrate SFPFR's superior performance (PSNR: 27.74, FID: 37.13), with ablation studies validating each module's effectiveness. The proposed method advances both technical boundaries and clinical applications, having potential applications for the evaluation, diagnosis, and treatment planning of facial paralysis.

Acknowledgments. The research in this paper was funded by the NSFC (32301702) and the Open Project of Shandong Provincial Engineering Research Center for Intelligent Service Assurance of Low-altitude Flight (KF2024SD008).

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Junsik Kim, Hyungwha Jeong, Jeongmok Cho, Changsik Pak, Tae Suk Oh, Joon Pio Hong, Soonchul Kwon, and Jisang Yoo. Numerical approach to facial palsy using a novel registration method with 3d facial landmark. *Sensors*, 22(17):6636, 2022.
2. Xiangyu Zhu, Chang Yu, Di Huang, Zhen Lei, Hao Wang, and Stan Z Li. Beyond 3dmm: Learning to capture high-fidelity 3d face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1442–1457, 2022.
3. Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *2021 international conference on 3D Vision (3DV)*, pages 453–463. IEEE, 2021.
4. Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023.
5. Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.
6. Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.
7. Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

8. John P Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H Pighin, and Zhigang Deng. Practice and theory of blendshape facial models. *Eurographics (State of the Art Reports)*, 1(8):2, 2014.
9. George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2490–2501, 2024.
10. William AP Smith. The perspective face shape ambiguity. In *Perspectives in shape analysis*, pages 299–319. Springer, 2016.
11. Bernhard Egger. *Semantic morphable models.* PhD thesis, University_of_Basel, 2017.
12. Yajing Chen, Fanzi Wu, Zeyu Wang, Yibing Song, Yonggen Ling, and Linchao Bao. Self-supervised learning of detailed 3d face reconstruction. *IEEE Transactions on Image Processing*, 29:8696–8705, 2020.
13. Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 959–968, 2019.
14. Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.
15. Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pages 694–711. Springer, 2016.
16. Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
17. Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 601–610, 2020.
18. Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
19. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
20. Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 285–295, 2019.
21. Zidu Wang, Xiangyu Zhu, Tianshuo Zhang, Baiqin Wang, and Zhen Lei. 3d face reconstruction with the geometric guidance of facial part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1672–1682, 2024.