# SUPPLEMENTARY MATERIALS: DETAILED 3D FACE RECONSTRUCTION IN LARGE POSE SCENARIOS

*Xinyu Li*[*], *Xitie Zhang*[*], *Suping Wu*[†] , *Ruijie Peng, JKehua Ma, Xiang Zhang*

School of Information Engineering, Ningxia University, Yinchuan, China
pswuu@nxu.edu.cn

## 1. DETAILS RECONSTRUCTION IN LARGE POSE SCENARIOS

### 1.1. Problem Analysis and Motivation

*1.1.1. Existing methods struggle to reconstruct detailed features of non-visible areas in faces with large poses.*

Existing detailed 3D face reconstruction methods typically employ the pipeline illustrated in Figure 1, utilizing a network that estimates a depth displacement map from the input image and a coarse 3D facial model. This map, representing detailed facial features, is then transformed into a 3D mesh and integrated onto the surface of the coarse 3D face to create a detailed 3D face. However, when the input image is under a large pose, the details of the non-visible areas are lost, making it challenging for the network to accurately estimate details in these regions. Consequently, this results in disordered or missing details in the non-visible areas of the 3D face, as shown in Figure 2.

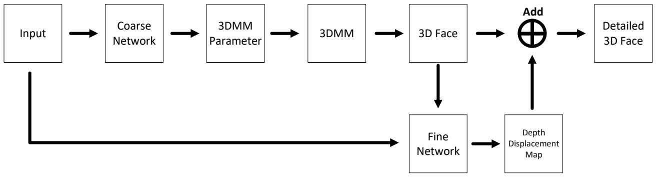

**Fig. 1**. The pipeline of existing detailed 3D face reconstruction methods.

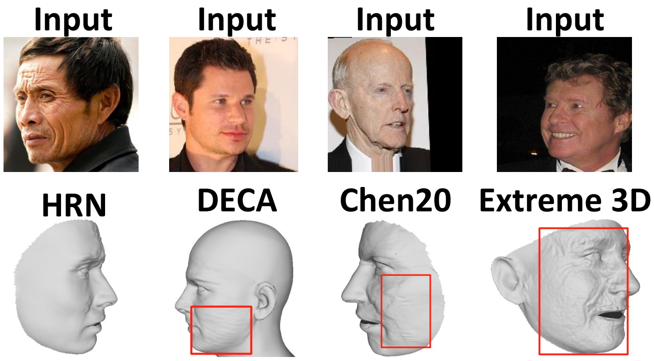

**Fig. 2**. In the first image, the visible parts of the face exhibit a rich array of wrinkles. Consequently, it is reasonable to assume that the non-visible parts should also possess some degree of wrinkling. However, the results from the HRN[1] show no wrinkles at all in these areas. For the other images, the corresponding results demonstrate a disarray of details.

*1.1.2. The solutions offered by GAN or diffusion models are suboptimal*

we attempted to utilize these models to first infer the details of non-visible areas and then proceed with reconstruction, aiming to achieve a detailed 3D face. We selected a state-of-the-art (SOTA) facial UV repair model, ”OSTeC[2]”, two general repair diffusion models, ”RePaint[3]” and ”DiffPIR[4]”, and a generative adversarial model ”SPI[5]” that directly generates 3D faces from image (which could generate images and videos but does not produce 3D mesh). Following this, we adopted Chen20[6] as our reconstruction pipeline, integrating the generative or diffusion models into this pipeline as illustrated in Figure 3.

We present the results of this approach in Figure 4, demonstrating that while OSTeC repairs the most detail, it often generates incorrect content, such as the unintended addition of glasses in the second and third images, rendering the inference results unreliable. Furthermore, OSTeC suffers from failures due to its inability to detect key points, as shown in the fourth image. The other two diffusion model approaches produce disorganized inference content, as seen in first, third, and fourth image, which leads to disordered details in the reconstructed 3D faces.

SPI[5] is capable of directly generating facial images that carry 3D information. In the first image of Figure 5, it produces credible details for the non-visible areas; however, in the other three images, the shape of the 3D faces is severely distorted. Moreover, SPI does not directly generate a 3D mesh, which is inconvenient for applications involving 3D faces. Therefore, GAN and diffusion models do not ideally resolve the issue of existing methods struggling to reconstruct

[*] represents equal contributions and [†] represents corresponding author.

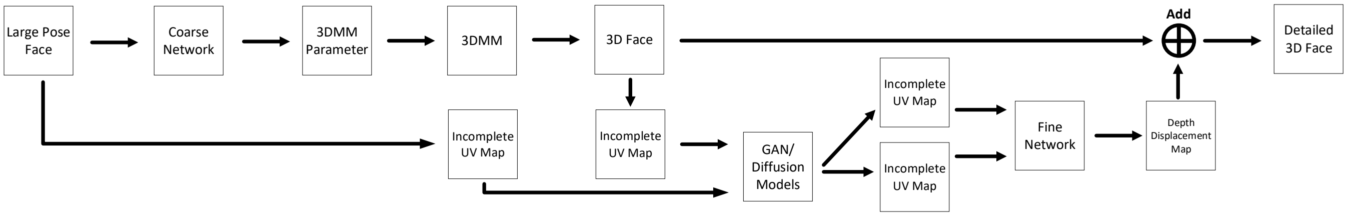

**Fig. 3**. Pipeline to addressing the problem of existing methods using GAN or diffusion models.

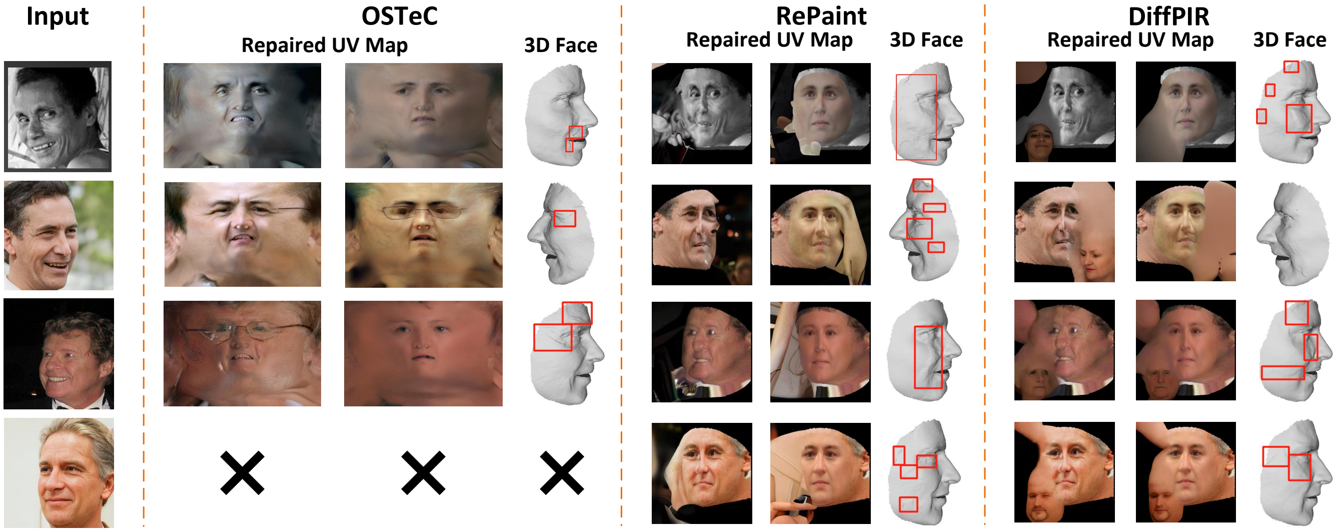

**Fig. 4**. UV maps repaired using GAN or diffusion models, and the 3D faces reconstructed using these repaired UV maps.

details in non-visible areas. Due to their non-end-to-end architecture, integrating them into existing methods incurs additional overhead. The average GPU memory utilization for the four models we applied is 33,160 MB, and the average inference time is 2,196 seconds per image. This also makes widespread adoption challenging in terms of automation and ease of use.

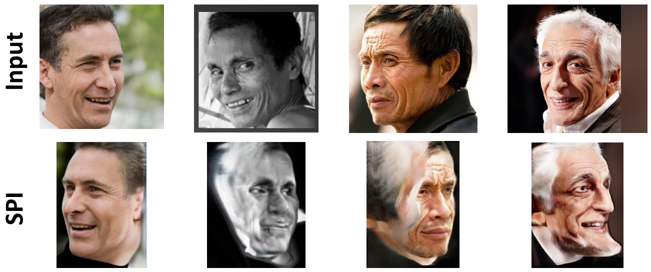

**Fig. 5**. Display of results from the GAN method SPI[5] that directly generates 3D faces. This method could only output images and videos, without any 3D mesh.

### 1.2. Loss Function

**Coarse 3D Face Reconstruction.** The total loss function is denoted as:

$$\mathcal{L} = w_1\mathcal{L}_p + w_2\mathcal{L}_{lm} + w_3\mathcal{L}_{id} + w_4\mathcal{R}_{param} \quad (1)$$

where $\mathcal{L}_p$ is the pixel loss and it calculates the average $L_{2,1}$-distances between the value of each pixel of the input image and 2D image which is rendered by the coarse 3D face. $\mathcal{L}_{lm}$ is the landmark consistency loss, which measures the average $L_2$-distance between the GT 68 2D landmarks of input image and the 68 landmarks of the reconstructed 3D face. $\mathcal{L}_{id}$ is the perceptual identity loss and calculates the $L_2$-distance between the two feature vectors extracted by the same VGG-16 for input image and the image rendered by coarse 3D face. In addition, $\mathcal{R}_{param}$ represents the regularization term for 3DMM parameters estimated by VGG-16. The weights $w_1$, $w_2$, $w_3$, and $w_4$ are constant values to balance the influence of each loss term. Specific objective function details are in the supplementary material.

**Detailed 3D Face Reconstruction.** We adopt the same objective functions to train DFCN in both small and large pose reconstructions. The total objective function is denoted as:

$$\mathcal{L} = a_1\mathcal{L}_p + a_2\mathcal{L}_s + a_3\mathcal{R}_{disp} \quad (2)$$

where $\mathcal{L}_p$ is the pixel loss. $\mathcal{L}_s$ is the smoothness loss which ensures the smooth displacement depth change on faces. In addition, $\mathcal{R}_{disp}$ is a regularization term for smoothness loss.

$$\begin{aligned}\mathcal{L}_s = \textstyle\sum_{i \in \mathcal{V}_{UV}} \sum_{j \in \mathcal{N}(i)} b_1 \|\Delta n(i) - \Delta n(j)\|^2 \\ + b_2 \|\Delta z(i) - \Delta z(j)\|^2\end{aligned} \quad (3)$$

where $\mathcal{V}_{UV}$ are vertices in the UV space and $\mathcal{N}(i)$ is the neighborhood of vertex i with a radius of 1. $\Delta n(i)$ and $\Delta n(j)$ represent the difference of UV normal map before and after adding displacement depth map for $i$ and $j$, respectively. $\Delta z(i)$ and $\Delta z(j)$ represent the difference in displacement depth map before and after adding UV normal map for $i$ and $j$, respectively. The weights $a_1$, $a_2$, $a_3$, $b_1$, and $b_2$ are constant values to balance the influence of each loss term.

### 1.3. Experiment Setting

We utilized the GitHub project of Chen20[6] as our baseline, adopting its network architecture and hyperparameters while incorporating our proposed self-supervised and RGB to Depth methods. To preserve the network's ability to reconstruct small-pose facial details, we employed a network pretrained on Chen20 for 20,000 steps as our inference network and further trained it for 5,000 steps. The training was conducted with a batch size of 10, and all other hyperparameters remained consistent with Chen20.

Given that our research focuses on the 3D mesh details of large pose face reconstructions, which are integrated with coarse 3D facial models, there is currently no benchmark available that isolates the influence of the coarse 3D face to solely evaluate the detail in large pose scenarios. Therefore, we assess our method through qualitative experiments.

## 2. COARSE 3D FACE RECONSTRUCTION

### 2.1. Experimental Setup

We employed the VGG-16, pre-trained on the VGG-Face[7] and Celeb A[8] datasets, as our reconstruction network. We controlled the pose parameters to reconstruct 3D faces in various poses for each image in a Celeb A subset, and then rendered these faces into images. To explore the significance of different poses, we separately used different pose rendered images to train the pre-trained VGG-16. After training, we quantitatively assessed these models on the NOW validation[9] dataset using the protocol of [10], with introduction of the dataset and protocol provided in the section 2.3.2. We hypothesize that the better the quantitative results of a VGG-16 model, the more beneficial its used pose parameters are for training.

### 2.2. Impact of The Proportion of Large Poses on Coarse Reconstruction

#### *2.2.1. Large Pose is More Important*

Following the setup of 3DDFA [11], angles between $[0, 45]$ are categorized as small to medium poses, while angles between $[45, 90]$ are considered large poses. We selected yaw and pitch angles from the set -15, 0, +15 to generate 8 small poses. For each input image, we render 8 images using these small poses, with examples shown in Figure 12(a). Similarly, by choosing values from the set -60, 0, +60, we render 8 large pose images for each input image, as depicted in Figure 12(b). We trained the pre-trained network separately with these two sets of images and performed quantitative evaluations, shown in Table 1. The results indicate that the 3D NME of the pretrained network is 3.559. Training with large pose images results in a lower 3D NME compared to training with small pose images, verifying that large poses are more beneficial for training reconstruction networks.

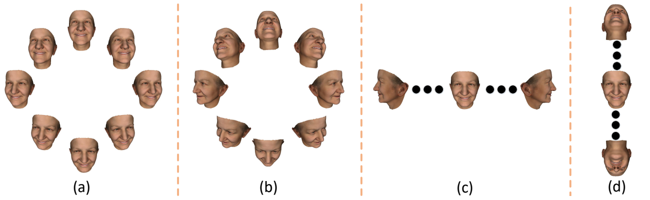

**Fig. 6**. Rendered Images Generated from Controlled Pose Parameters. (a), (b), (c), and (d) represent large poses, small poses, yaw angles, and pitch angles, respectively.

**Table 1**. Quantitative Results from Training Sets with Different Poses. 'Pre-trained' refers to pre-trained reconstruction network. 'Small Pose' and 'Large Pose' respectively denote the continuation of training the pre-trained network with images of small and large poses.

| Model | 3D NME |
|---|---|
| Pre-trained | 3.559 |
| Small Pose | 2.766 |
| Large Pose | 2.593 |

#### *2.2.2. Yaw Angle is More Important*

Although we have established the significance of large poses, there are three fundamental orientation directions in 3D space: yaw, pitch, and roll. Therefore, we need to further determine which direction is more critical. Since roll does not significantly alter the 3D structural information of the face, we primarily investigated the effects of yaw and pitch angles. To do this, we uniformly sampled 9 yaw angles within the range of [-90, +90] and rendered 9 images for each input image, as shown in Figure 12(c). These images encompass a full range of yaw angles but maintain a pitch angle of 0,

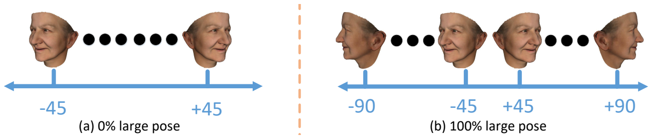

**Fig. 7**. We rendered 11 sets of images with increments of 10%, where the proportion of large poses ranged from 0% to 100% in steps of 10%. The figures display the rendered images at 0% and 100% large pose proportions, while images for other proportions are not shown.

thus containing only information about the yaw without any pitch information. Similarly, we produced images that only incorporate pitch information without yaw, as depicted in Figure 12(d). We trained the pre-trained reconstruction network separately with these two sets of images and performed quantitative evaluations, with the results presented in Table 2. The findings indicate that networks trained with images varying in yaw angle exhibited the lowest reconstruction errors, demonstrating that the yaw angle is more critical. Combining this with the earlier insight that larger poses significantly aid network training, we can confirm that large poses in the yaw direction are the most beneficial for training reconstruction networks.

**Table 2**. Quantitative Results from Training Sets with Different angles. Training with datasets characterized by yaw angle results in lower reconstruction errors.

| Model | 3D NME |
|---|---|
| Pre-trained | 3.559 |
| Pitch Angle | 2.500 |
| Yaw Angle | 2.400 |

**Table 3**. We consolidated four datasets—Celeb A, VGGFace2, BUPT-Balancedface, and VoxCeleb2—and analyzed the proportion of poses within them. It was observed that large pose images constitute only 6.342% of the total.

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

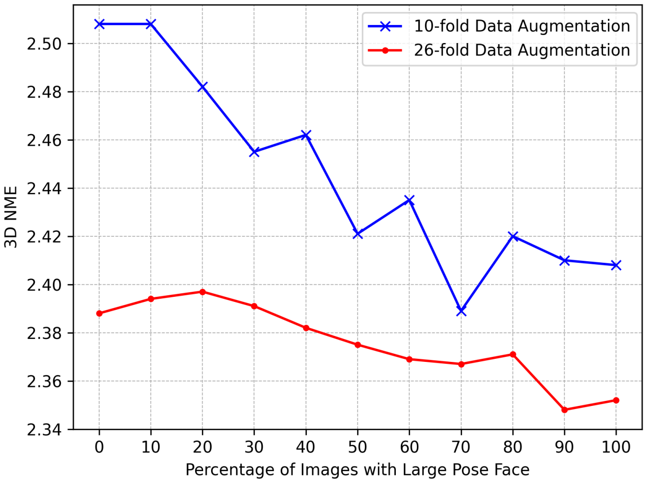

**Fig. 8**. We conducted training and quantification using different proportions of large-pose images across two training scales. The results indicate a downward trend in reconstruction error as the proportion of large poses increases. This suggests that the current 6.342% proportion of large-pose training images in existing methods is not conducive to optimal network learning.

### 2.2.3. *Low Proportion of Large Pose Faces in Existing Training Sets*

Existing detailed 3D face reconstruction methods usually utilize datasets such as Celeb A[8], VGGFace2[7], BUPT-Balancedface[12], and VoxCeleb2[13] for training. We analyzed the number and proportion of large pose faces within these datasets based on their yaw angles, with the results displayed in Table 3. It is evident that large poses constitute only a small fraction of the datasets. To investigate the impact of the proportion of large poses on reconstruction, we conducted further verification. Specifically, we uniformly sampled ten poses within a yaw angle range of [-45, +45] and rendered 10 images with small poses for each input image, as illustrated in Figure 13(a). This resulted in a dataset with 0% large poses. We then increased the yaw angle of all images by 9 degrees incrementally, thereby increasing the proportion of large poses by 10% with each increment. This process is continued until the large pose ratio reaches 100%, as shown in Figure 13(b). Thus, we obtain 11 different sets of images with large pose ratios ranging from 0%, 10%, ..., to 100%. We trained the reconstruction network with each set of images and conducted quantitative evaluations. To determine if the same pattern holds for larger training volume, we rendered 26 images to perform the same experiment, and the results are displayed in Figure 14. Both large and small dataset sizes showed a trend where overall error decreased as the proportion of large poses increased. This suggests that the 6.342% proportion of large poses in existing methods is highly detrimental to 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 recommend increasing the proportion of large pose images in training datasets to enhance the accuracy of existing methods.

## 2.3. Experiment

### 2.3.1. Baseline

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[1], but since its training code is not publicly available, we selected the latest SOTA models with available training code, "DECA"[14] and "Chen20"[6], which are also highly influential. We use their GitHub projects as baselines for our experiments and retained the hyperparameter settings provided in these projects.

### 2.3.2. Experiment Setting

For DECA, we experimented with various large pose proportions and determined the optimal setting to render 22 images, with the pitch and roll angles set to 0, and yaw angles sampled every 2.5 degrees within the ranges of [-65, -90] and [65, 90], achieving a 100% large pose proportion. Additionally, the batch size was set to 3, with a total of 6200 training steps. For Chen20, the optimal configuration selected involves rendering 26 images with the pitch and roll angles set to 0, and yaw angles sampled every 2.5 degrees within the ranges of [-60, -90] and [60, 90], totaling 26 different angles. The batch size for this setup is 5, with a training duration of 310 steps.

We performed quantitative assessments using the MICC Florence [15], NOW validation[9], and 3dMDLab[16] (including 3dr and 3ds subsets) datasets. To independently study the results under small and large pose conditions, we divided each dataset into small-pose or large pose images for separate evaluations. A brief description of each dataset is as follows.

- MICC Florence. In MICC, videos are taken on 53 subjects under Indoor Cooperative, PTZ Indoor, and PTZ Outdoor conditions. The ground truth 3D scans are provided for 52 out of the 53 people. We select the Indoor Cooperative condition which has the clearest face, and crop a face picture with a small pose and a face picture with a large pose for each available subject from the clear video frame as the evaluation dataset.

- NOW Validation. In NOW Validation, there are 20 subjects in total, each subject has face pictures from 5 or 6 poses and a ground truth 3D scan. For each subject, we select a face picture with small pose and a face picture with large pose as the evaluation dataset under the two poses, respectively.

- 3dMDLab. It is a high-quality large pose dataset that includes two subsets: 3dMDLab-real (3dr) and 3dMDLab-synthetic (3ds). The former consists of 8 real large pose facial images and their corresponding ground truth 3D scans. The latter comprises six synthetic large pose facial images and their corresponding ground truth 3D scans. Additionally, all images are high-resolution (2048x2448 pixels) with a true color range (24 bits per pixel).

We employed the widely adopted protocol[10] to calculate the 3D normalized mean error (NME) between each point on the reconstructed 3D face and the ground truth 3D face. This error metric serves as the primary evaluation criterion.

## 3. REFERENCES

[1] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie, "A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 394–403.

[2] Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou, "Ostec: One-shot texture completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 7628–7638.

[3] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and L Repaint Van Gool, "Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471.

[4] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool, "Denoising diffusion models for plug-and-play image restoration," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (NTIRE)*, 2023.

[5] Fei Yin, Yong Zhang, Xuan Wang, Tengfei Wang, Xiaoyu Li, Yuan Gong, Yanbo Fan, Xiaodong Cun, Ying Shan, Cengiz Oztireli, and Yujiu Yang, "3d gan inversion with facial symmetry prior," *arXiv preprint arXiv:2211.16927*, 2022.

[6] 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*, vol. 29, pp. 8696–8705, 2020.

[7] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[8] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[9] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black, “Learning to regress 3D face shape and expression from an image without 3D supervision,” in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 7763–7772.

[10] Z. H. Feng, P. Huber, J. Kittler, P. J. Hancock, X. J. Wu, Q. Zhao, P. Koppen, and M. Rtsch, “Evaluation of dense 3d reconstruction from 2d face images in the wild,” in *arXiv*, 2018.

[11] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li, “Face alignment across large poses: A 3d solution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[12] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang, “Racial faces in the wild: Reducing racial bias by information maximization adaptation network,” in *Proceedings of the ieee/cvf international conference on computer vision*, 2019, pp. 692–702.

[13] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.

[14] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart, “Learning an animatable detailed 3d face model from in-the-wild images,” *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–13, 2021.

[15] A. D. Bagdanov, A Del Bimbo, and I. Masi, “The florence 2d/3d hybrid face dataset,” *ACM*, p. 79, 2011.

[16] James Booth, Anastasios Roussos, Evangelos Ververas, Epameinondas Antonakos, Stylianos Ploumpis, Yannis Panagakis, and Stefanos Zafeiriou, “3d reconstruction of “in-the-wild” faces in images and videos,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2638–2652, 2018.