

In this section we reveal more implementation details and provide more analytical and visual comparisons on the shape reconstruction and segmentation performances.

A. Implementation Details

A.1. The neighbor loss

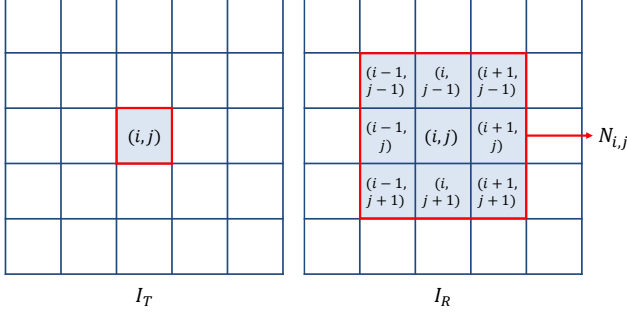


Figure 1. Visual explanation of the neighbor loss (Eq. (1)). For a pixel at (i, j) on the target image I_T (left), we compare it with its neighboring region, $N_{i,j}$, on the reconstructed image I_R (right).

In this paper, we introduce a new image-level neighbor loss, L_{nbr} , that compares one pixel in the target image to a small region in the reconstructed image:

$$L_{nbr} = \sum_{x \in \Omega} \left\| \min_{x' \in N(x)} \|I_T(x) - I_R(x')\|_2 \right\|_2^2 \quad (1)$$

As shown in Fig. 1, for every pixel $I_{T(i,j)}$ in the target image, we search in a 3×3 neighborhood $N(i, j)$ in the reconstructed image I_R for the pixel that is most similar to $I_T(i, j)$ in intensity. This neighbour loss accounts for small misalignments of the face model during segmentation.

A.2. Training details

To train our proposed pipeline, the Adadelta optimizer is used, with an initial learning rate of 0.1, and a decay rate of 0.99 at every 5k iterations. The learning rate for the segmentation network is 0.06 times the one for the reconstruction network. In every 30k iterations, 25k iters are for the face autoencoder training, and the rest are for training the segmentation network. For initialization, the face autoencoder is trained for 300k iterations. Afterwards, the face autoencoder and segmentation network are trained jointly for 200k iterations. The speed is evaluated on an RTX 2080 Ti, with batch size 12. It takes about 120 hours for the initialization of the face autoencoder, and about 80 hours to train the complete pipeline. After the training, it takes 49 ms for reconstruction and 70 μ s for segmentation on average for one image. The reconstruction and the segmentation networks have 25.6M and 34.5M parameters, respectively.

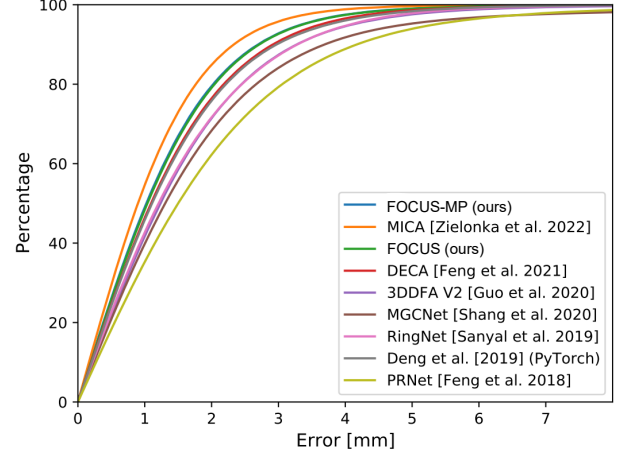


Figure 2. Quantitative comparison of the 3D reconstruction accuracy on the NoW [8] testset. The methods shown include: MICA [11], DECA [3], the work of Dib et al. [2], 3DDFA V2 [5], MGCNet [9], RingNet [8], Deep3D (pytorch version) [1], and PRNet [4]

B. Quantitative Analysis

B.1. Reconstruction Performance.

Fig. 2 shows the cumulative error curves of the proposed method and the state-of-the-arts regarding the NoW testset. With a higher percentage of sampling points with lower errors, FOCUS performs the best on the NoW testset.

We further compare analytically the distributions of reconstruction errors of DECA [3] and FOCUS on the NoW validation set, as shown in Fig. 3. To further disentangle the influence of outliers from other factors, we categorize the samples according to the yaw angles (rounded off to the nearest 10), and use the error bars under different poses to reflect the distribution of the reconstruction errors. It is obvious from the plots that FOCUS exceeds DECA in mean errors and yields in much lower variations, even without identity supervision (which emphasize the shape consistency of a same identity) and with significantly less training data. Besides, the lower gap between errors and deviations under occluded and unoccluded conditions shows that the proposed method improves the outlier robustness. The comparison between the FOCUS and FOCUS-MP pipelines also indicates that the misfit prior improves the overall reconstruction accuracy.

B.2. Segmentation Performance.

Tab. 1 shows the segmentation performance on the Celeb A HQ testset, which indicates that the masks predicted by our method show a competitive accuracy, precision, and F1 score, compared to the skin detector used in [1].

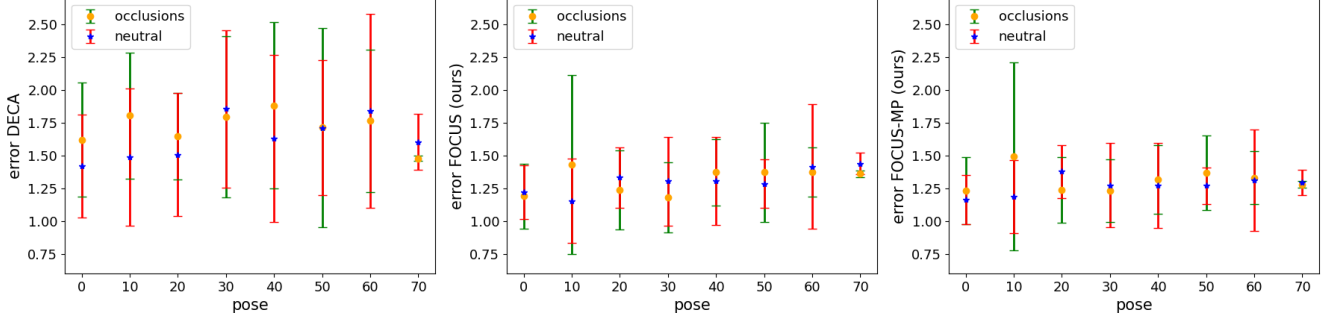


Figure 3. The distribution of the reconstruction errors on the neutral and occluded subsets of the full NoW validation set. The results of DECA [3] are on the left, our FOCUS model in the middle and our FOCUS-MP on the right. The x axis indicates the approximated poses of the samples (rounded off to the nearest 10), and the y axis denotes the reconstruction error.

Table 1. Evaluation of occlusion segmentation accuracy on the CelebA-HQ testsets.

Method	Unoccluded				Occluded				Overall			
	ACC	PPV	TPR	F1	ACC	PPV	TPR	F1	ACC	PPV	TPR	F1
Deep3D [1]	0.95	0.98	0.97	0.97 ± 0.06	0.84	0.86	0.96	0.90 ± 0.08	0.89	0.92	0.96	0.93 ± 0.07
FOCUS (ours)	0.92	0.99	0.93	0.96 ± 0.02	0.86	0.95	0.87	0.91 ± 0.06	0.89	0.97	0.90	0.93 ± 0.05

Table 2. Reconstruction error (mm) on NoW validation subsets.

Method	Unoccluded Subset			Occluded Subset		
	median	mean	std	median	mean	std
Backbone-Supervised	1.02	1.25	1.04	1.05	1.29	1.09
Backbone-Supervised-MP	1.00	1.23	1.02	1.03	1.28	1.08
Backbone-cutmix	1.05	1.28	1.04	1.08	1.33	1.11
Backbone-cutmix-MP	1.04	1.27	1.03	1.08	1.32	1.10
Backbone-cutout	1.03	1.28	1.06	1.09	1.34	1.10
Backbone-cutout-MP	1.02	1.25	1.04	1.08	1.32	1.09

B.3. Significance of the Misfit Prior

Regarding Table 2 in the paper, a paired t-test on the NoW validation set shows a two-sided p-value of $3.4\text{e-}19$, less than 0.05. Hence, the mean errors before and after using the prior are not equal, indicating that the misfit prior brings a substantial increase in reconstruction accuracy.

Tab. 2 shows the misfit prior generalizes well for our fully-supervised counterparts.

B.4. Hyper-parameter Analysis

In this section we systematically evaluate the influence of the hyper-parameters, η_1 to η_5 , used for segmentation. The total loss for training the segmentation network is: $L_S = \eta_1 L_{nbr} + \eta_2 L_{dist} + \eta_3 L_{area} + \eta_4 L_{presv} + \eta_5 L_{bin}$, with $\eta_1 = 15$, $\eta_2 = 3$, $\eta_3 = 0.5$, and $\eta_4 = 2.5$, and $\eta_5 = 10$. We call this set of parameters as ‘standard parameters’. We use the control variates method, namely changing one of the parameters while fixing the others, to compare the influence of each hyper-parameters. The accuracy (ACC), precision (Positive Predictive Value, PPV), recall rate (True

Positive Rate, TPR), and F1 score (F1) reflect the segmentation performance. We use the AR dataset [7] because the segmentation labels are more accurate.

As shown in Fig. 4, with the increase of the neighbour loss L_{nbr} or the perceptual-level loss L_{dist} , more pixels are segmented as non-facial. On the contrary, when the area loss L_{area} or the pixel-wise preserve loss L_{presv} increases, more pixels are taken as face. This observation is consistent with our theory in section 3.2. Fig. 4 also indicates that the indices are positively related to the area loss L_{area} and preserving loss L_{presv} , and are negatively related to the neighbour loss L_{nbr} and the perceptual-level loss L_{dist} . The binary loss, L_{bin} , barely affects the segmentation.

Note that we did not excessively tune the loss weights, therefore we expect that better settings exist which achieve even higher performance.

C. Qualitative Comparison and Ablation Study

In this section, we provide more visual results of our method on the Celeb A HQ testset [6], the AR dataset [7], and the NoW testset [8].

Figs. 5 to 8 show the results under general occlusions, extreme lighting, skin-colored occlusions, and large poses, respectively.

Figs. 9 and 10 provide a visual comparison among the ablated pipelines. It highlights that the outlier robust function is not robust to illumination variations, and the segmentation network brings great benefit to the robustness to illumination. The neighbour loss encourages the network to

produce smoother results, and the perceptual losses help to locate the occlusions more accurately. Generally, the reconstruction performance of our proposed method are the best one and the segmentation accuracies are also competitive.

Fig. 11 show the intermediate results during the EM-like training introduced in section 3.2. The estimated masks get more accurate given better reconstruction results and the reconstructed faces show more details when provided with better segmentation masks, indicating the synergy effect of the reconstruction and segmentation networks.

D. Societal Impact

In general, our FOCUS pipeline has the potential to bring face reconstruction to the real world and to save costs of occlusion or skin labeling, which is generally required in many existing deep-learning-based methods. The model-based reconstruction methods improved by our method could contribute to many applications, including Augmented Reality (AR), Virtual Reality (VR), surveillance, 3D design, and so on. Each of these applications may bring societal and economic benefits, and risks at the same time: The application of AR or VR could bring profits to the entertainment industry and also may result in unethical practices such as demonizing the image of others, identity fraud, and so on. The application of surveillance could help arrest criminals, yet might also invade the privacy and safety of others. The application in 3D design enables the quick capture of the 3D shape of an existing face but might also cause problems in portrait rights.

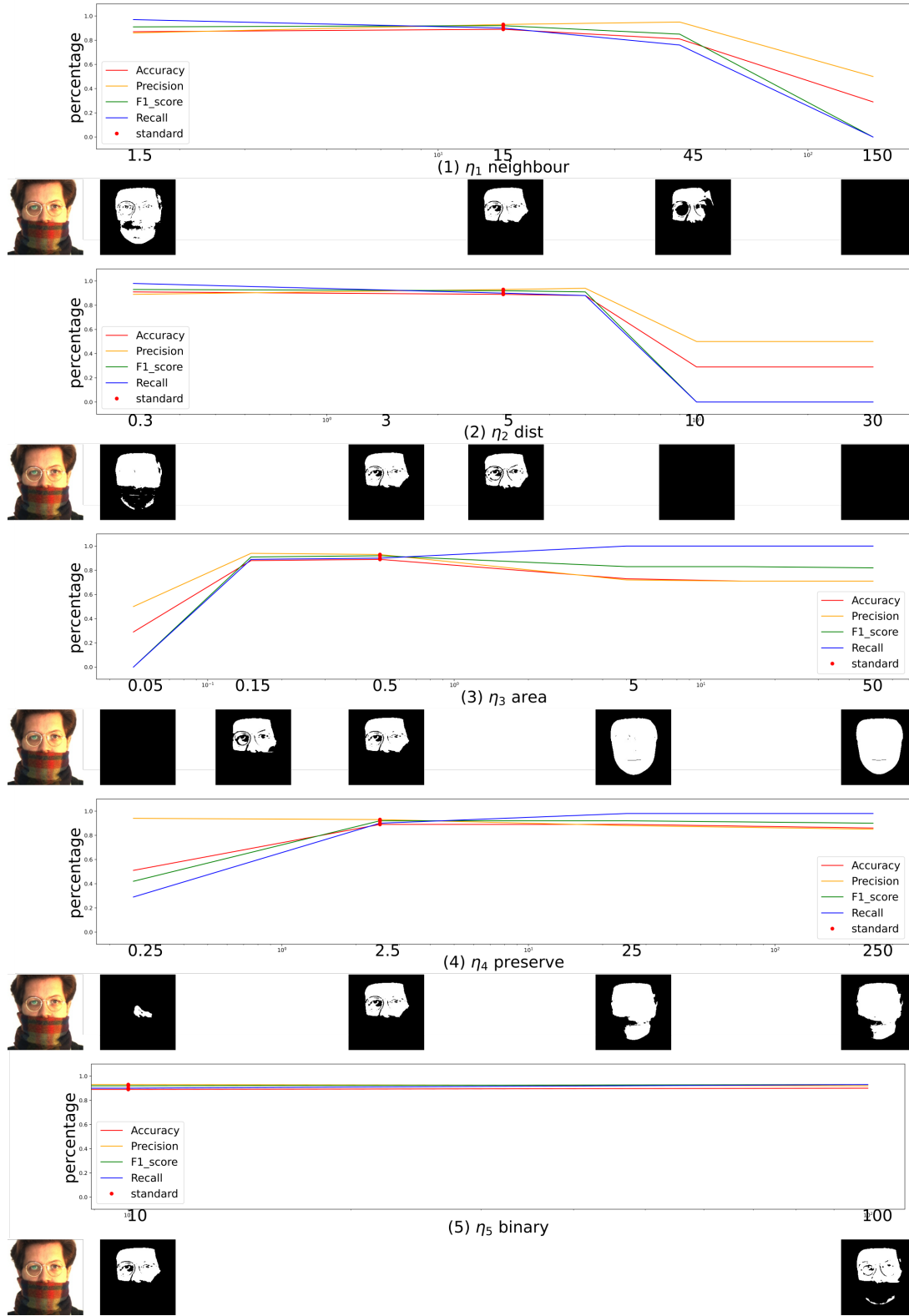


Figure 4. Analysis of hyper-parameters. The subplots show the change of for indices, namely accuracy, precision, F1 score, and recall rate, with the change of the hyper-parameters. The corresponding segmentation results are shown below each subplot. In each subplot, to evaluate the effect of each hyper parameter η_i , the other hyper-parameters $\eta_j (j \neq i)$ are fixed. The red dots denote the 'standard parameters' used in the paper.



Figure 5. Comparison on **random** samples in the Celeb A HQ [6] testset and the AR dataset [7]. (a) Target image. (b) and (c) Reconstruction and segmentation results of the Deep3D network [1]. (d) Reconstructed result of the MoFA network [10]. (e) and (f) Reconstruction and segmentation results of ours.



Figure 6. Comparison on samples with **extreme illumination** conditions in the Celeb A HQ [6] and the AR [7] testsets. (a) Target image. (b) and (c) Reconstruction and segmentation results of the Deep3D network [1]. (d) Reconstructed result of the MoFA network [10]. (e) and (f) Reconstruction and segmentation results of ours.



Figure 7. Comparison on samples with outliers that the **skin detector in [1]** fails to locate in the Celeb A HQ testset [6]. (a) Target image. (b) and (c) Reconstruction and segmentation results of the Deep3D network [1]. (d) Reconstructed result of the MoFA network [10]. (e) and (f) Reconstruction and segmentation results of ours.



Figure 8. Comparison on samples with outliers and **large poses** in the NoW Database [8] shows that our method can effectively handle outliers even when there are large poses. (a) Target image. (b) and (c) Reconstruction and segmentation results of the Deep3D network [1]. (d) Reconstructed result of the MoFA network [10]. (e) and (f) Reconstruction and segmentation results of ours.

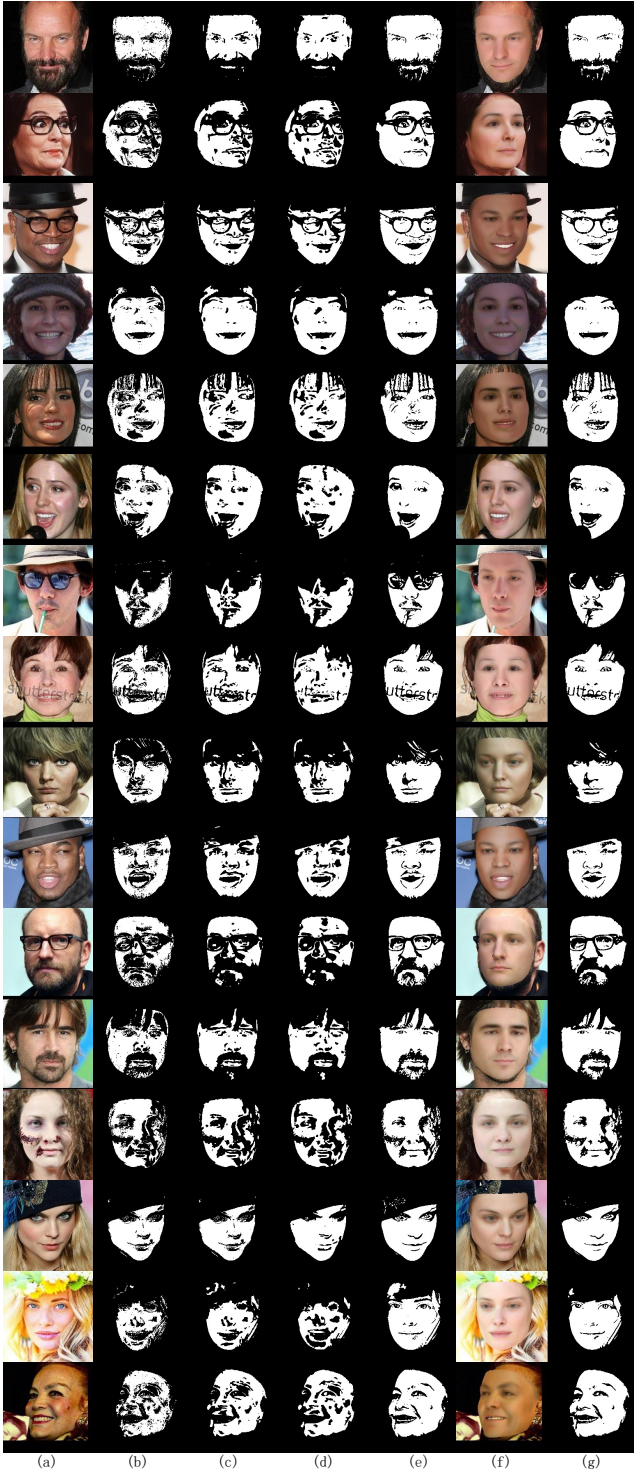


Figure 9. Qualitative comparison for ablation study on the Celeb A HQ testset [6]. From left to right are (a) target images, masks estimated by the (b) 'Pretrained', (c) 'Baseline', (d) 'Neighbour', and (e) 'Perceptual' pipelines, and (f) the reconstruction results and (g) predicted masks of FOCUS, respectively.

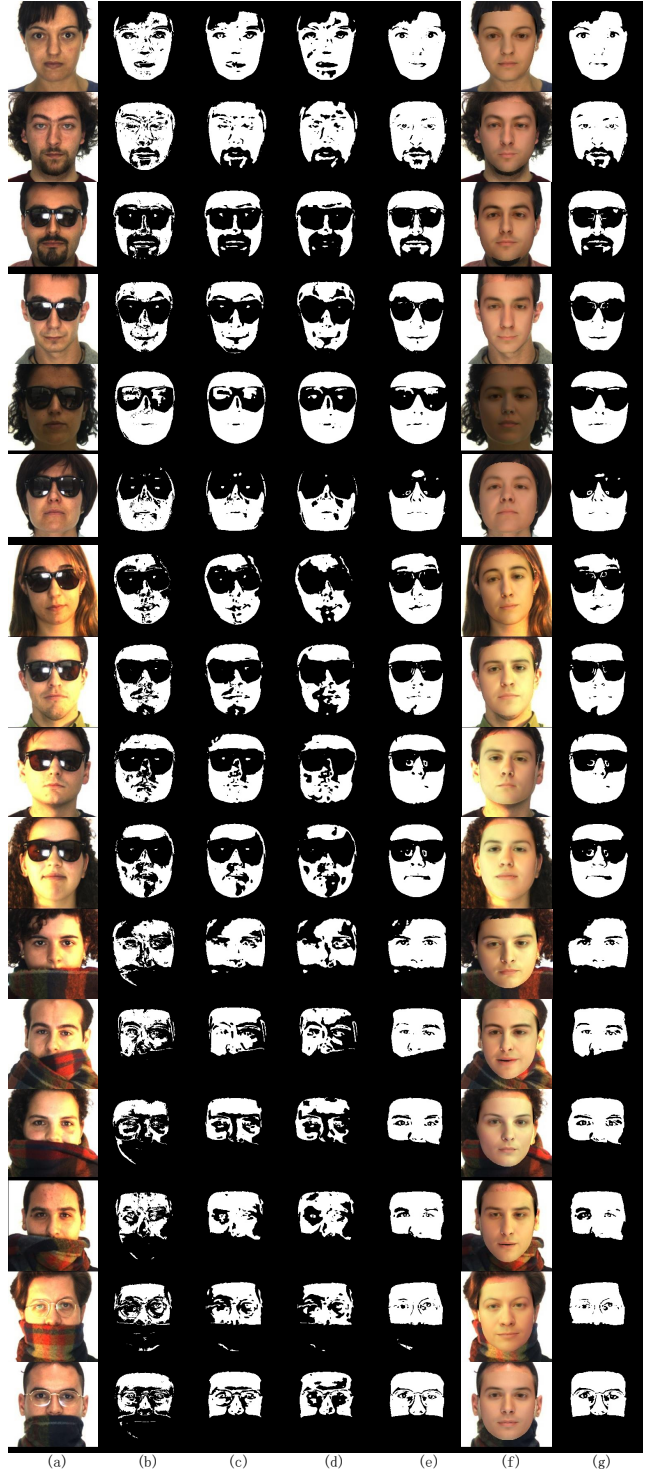


Figure 10. Qualitative comparison for ablation study on the AR testset [7]. From left to right are (a) target images, masks estimated by the (b) 'Pretrained', (c) 'Baseline', (d) 'Neighbour', and (e) 'Perceptual' pipelines, and (f) the reconstruction results and (g) predicted masks of FOCUS, respectively.

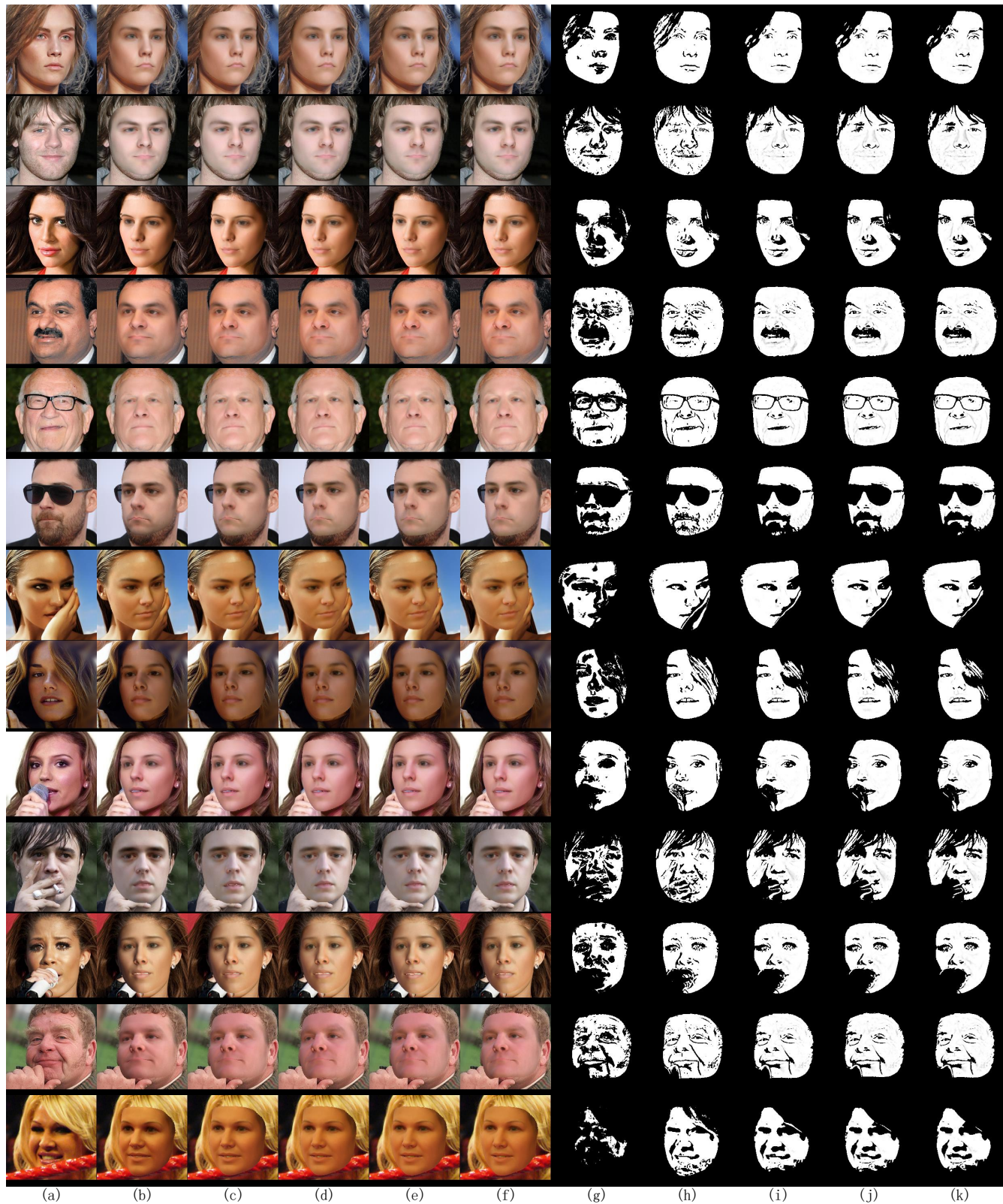


Figure 11. Target images (a) and intermediate results during the EM-like training. The intermediate masks and reconstructed faces predicted by: the initialized model introduced in section 3.3 (b and g), and the trained model after the first (c and h), second (d and i), third (e and j), and last (f and k) round of EM training.

References

- [1] 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 0–0, 2019. 1, 2, 5, 6, 7, 8
- [2] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 1
- [3] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *arXiv preprint arXiv:2012.04012*, 2020. 1, 2
- [4] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 1
- [5] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3D dense face alignment. In *European Conference on Computer Vision (ECCV)*, pages 152–168, 2020. 1
- [6] 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. 2, 5, 6, 7, 9
- [7] A. Martinez and Robert Benavente. The ar face database. *Tech. Rep. 24 CVC Technical Report*, 01 1998. 2, 5, 6, 9
- [8] 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. 1, 2, 8
- [9] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision (ECCV)*, volume 12360, pages 53–70, 2020. 1
- [10] Ayush Tewari, Michael Zollhofer, Hyeonwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 5, 6, 7, 8
- [11] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. *arXiv preprint arXiv:2204.06607*, 2022. 1