Supplementary Material for Paper: A Hierarchical Representation Network for Accurate and Detailed Face Reconstruction from In-The-Wild Images

Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, Xuansong Xie
DAMO Academy, Alibaba Group
{biwen.lbw, jianqiang.rjq, mengyang.fmy, miaomiao.cmm}@alibaba-inc.com,
xingtong.xxs@taobao.com

In this document, we present some additional statistics and more examples of the FaceHD-100 dataset in Sec. 1, ethics guidelines in Sec. 2, some implementation details in Sec. 3, head reconstruction as the extension work of our HRN (MV-HRN) in Sec. 4, more visualization results in Sec. 5, and discussions about the limitations of the proposed method and future work in Sec. 6.

1. The FaceHD-100 Dataset

This section provides more information about the FaceHD-100 dataset. The capturing subjects include 95% Asian, 3% white, and 2% black. Fig. 1 shows the age and gender distribution of the dataset, in which the ages of men and women are mainly concentrated between 17-35 years old, and the overall distribution is close to normal. In Fig. 2, we give an example of the 9-view face images from FaceHD-100, which shows the position distribution of the 9 cameras in our acquisition system. Fig. 3 presents some 9-view images and raw scans of different expressions from FaceHD-100. While capturing, each subject was asked to wear a hair covering to prevent hair from interfering.

We have signed an authorization agreement with each capturing subject, who grants us the exclusive rights to distribute, perform, and use the captured data within the scope of academic research (including for paper publication and representation) and legitimate business. And we will release the dataset for research purposes only.

2. Ethics Guidelines

In addition to FaceHD-100, other face datasets we use in the main paper and supplementary materials are licensed, granting us the right to use the data for research purposes, including publication of papers. Moreover, the face examples shown in the paper have also obtained the special authorization of the capturing subjects or followed the publishable list of the corresponding dataset.

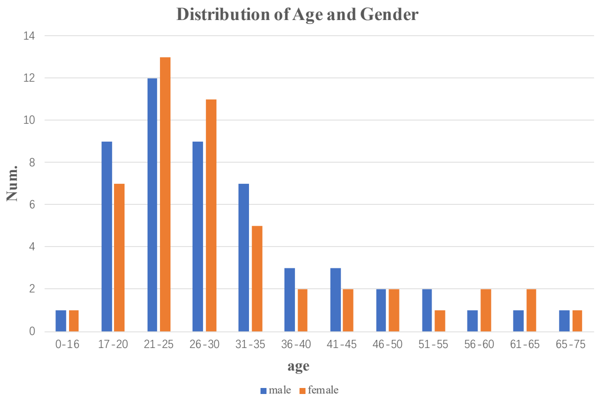


Figure 1. The age and gender distribution of the FaceHD-100 dataset.

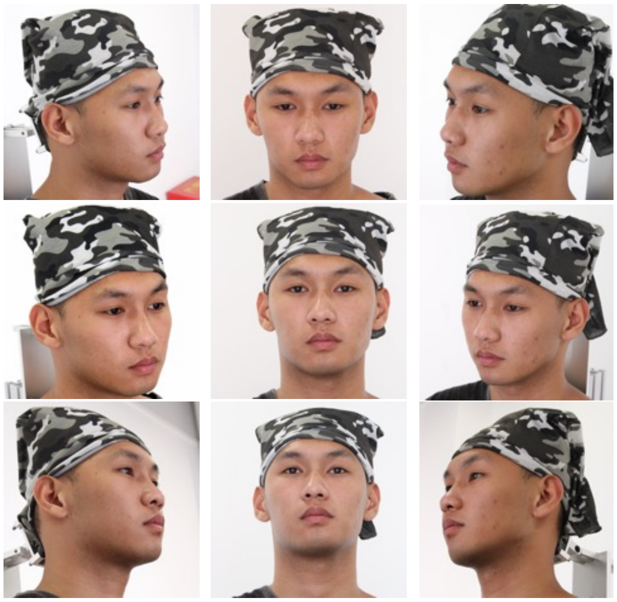


Figure 2. An example of the 9-view face images from the FaceHD-100 dataset.

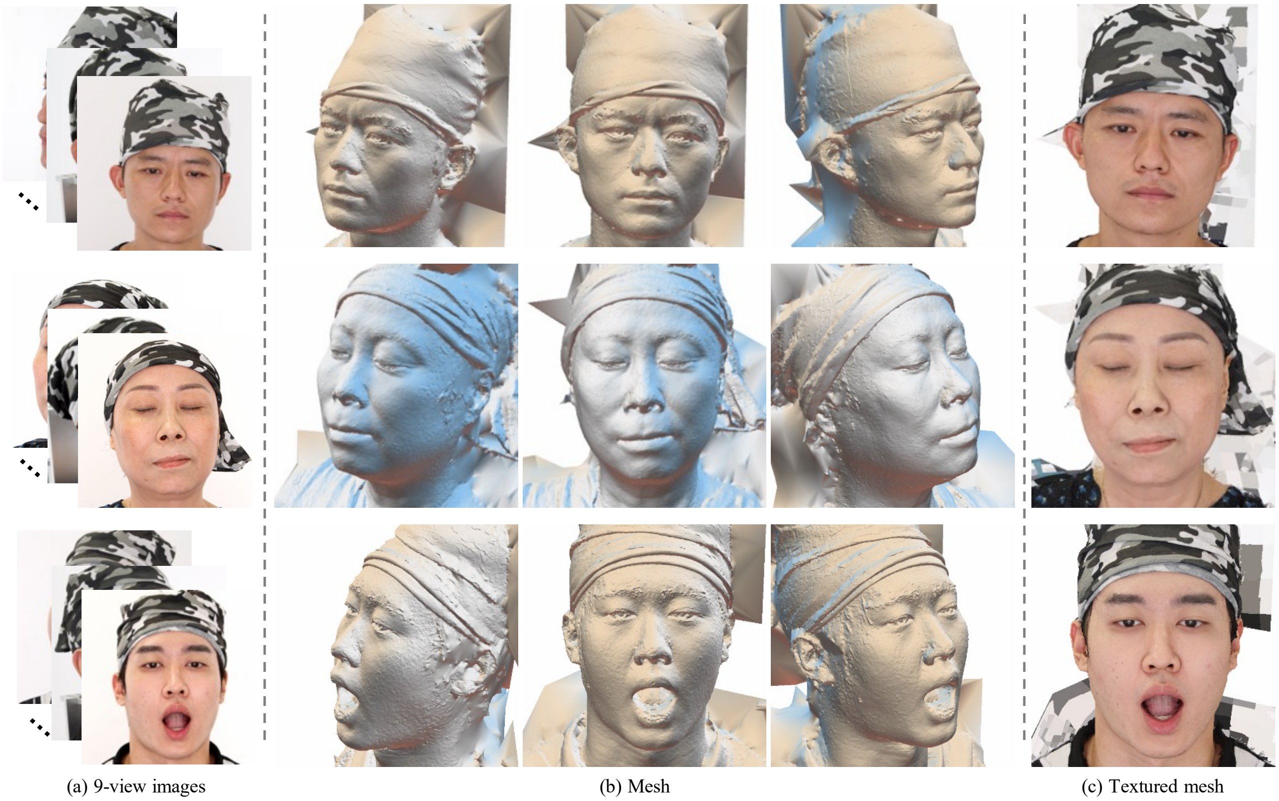


Figure 3. Some examples of different expressions from the FaceHD-100 dataset.

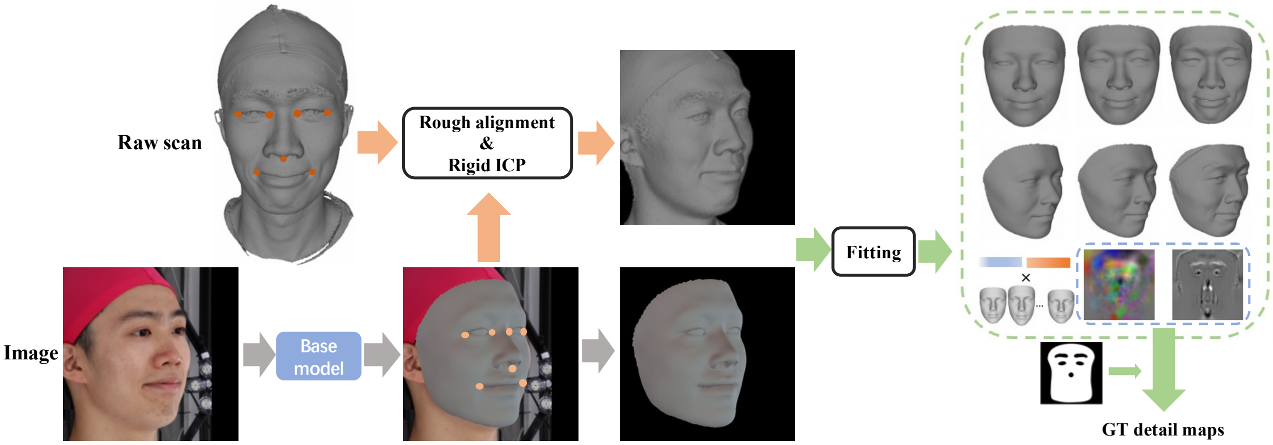


Figure 4. The pipeline of acquiring ground-truth deformation map and displacement map from a single image and the corresponding raw scan.

3. Implementation Details

3.1. Acquiring Ground-truth Detail Maps

As mentioned in Sec.3.3 in the main paper, to utilize the 3D data in our framework, we ought to transform the raw scan to align to the image in BFM space. Fig. 4 shows the pipeline of how we implement the transformation and acquire the ground-truth detail maps for training. Given a face image and its corresponding raw scan, we firstly employ the base model to predict a coarse mesh M_0 that is aligned to the image in BFM space. Then we obtain 7 landmarks from the raw scan and M_0 respectively to achieve rough alignment [13], and the rigid ICP [16] algorithm is further used to improve the alignment between the raw scan and M_0. Once we get the aligned scan, we are able to use the hierarchical representation to fit the scan as mentioned in Sec.3.3 in the main paper, and finally acquire the ground-truth deformation and displacement map for the input image (a mask is used in training to remove the noises of eyes, nose and hair area from raw scans). Note that since the base model is pre-trained in our network, we only optimize the detail maps

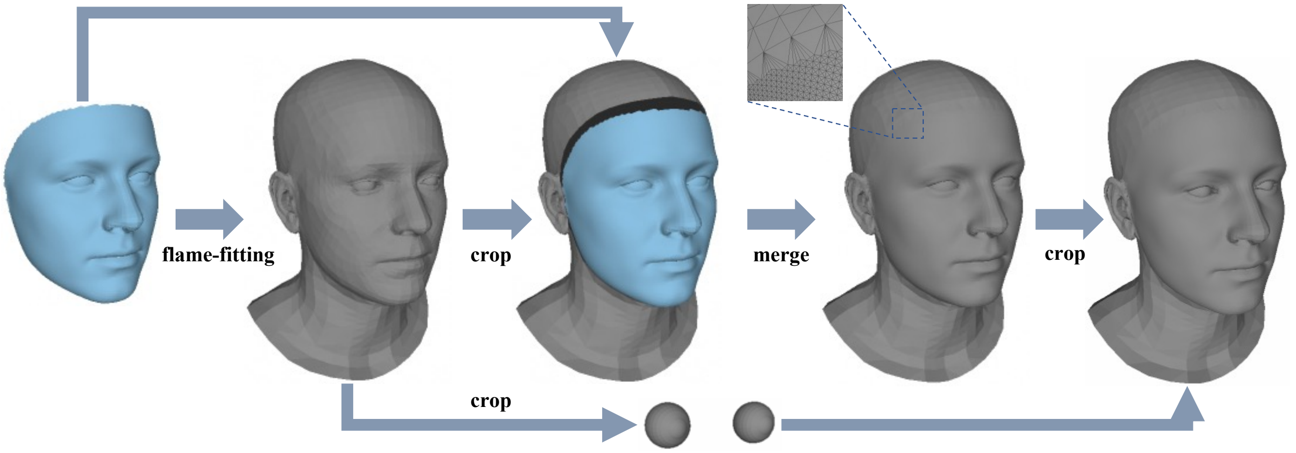


Figure 5. The pipeline of generating a new head model from the BFM model and FLAME model.

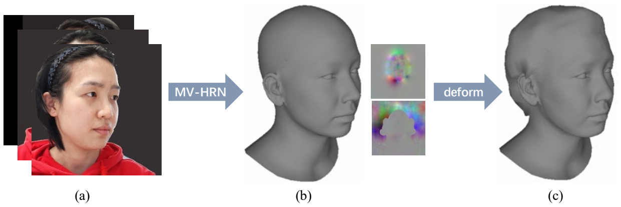


Figure 6. Simplified head reconstruction process. (a) Input multi-view images. (b) Predicted coarse head mesh, face, and hair deformation map. (c) deformed mesh.

and freeze the blendshape coefficients in the fitting process.

3.2. Loss Functions

As described in Sec.4.1 in the main paper, the training data is composed of two types of images: in-the-wild images and in-the-lab images. The former is used for training in a self-supervised manner, while the latter is combined with ground-truth detail maps generated following Sec. 3.1 and used for training in a supervised manner.

We use $\boldsymbol{R_1}$ and $\boldsymbol{R_2}$ to indicate the face rendered from $\boldsymbol{M_1}$ and $\boldsymbol{M_2}$ in Fig. 2 (main paper) respectively. Overall, the loss functions that we utilize for training consist of :
(i) two photometric losses [4] L_{photo_1} (between $\boldsymbol{I}$ and $\boldsymbol{R_1}$) and L_{photo_2} (between $\boldsymbol{I}$ and $\boldsymbol{R_2}$);
(ii) two perception-level losses [4] L_{per_1} (between $\boldsymbol{I}$ and $\boldsymbol{R_1}$) and L_{per_2} (between $\boldsymbol{I}$ and $\boldsymbol{R_2}$);
(iii) a landmark loss [4] L_{lan} (between $\boldsymbol{I}$ and $\boldsymbol{R_2}$);
(iv) a total variation loss [9] L_{tv} for the deformation map;
(v) an L1 regularization loss L_{reg} for the displacement map;
(vi) a contour-aware loss L_{con} between face mask and M_1;
(vii) two adversarial losses [8] L_{adv_mid} and L_{adv_high} for the deformation map and displacement map respectively;
(viii) an L1 loss L_{mid} for the deformation map and an L1 loss L_{high} for the displacement map in supervised training.

In summary, the joint loss for self-supervised and supervised training can be written as:

$$\begin{aligned}\mathcal{L}_{self} = &\lambda_1(L_{photo_1} + L_{photo_2}) + \lambda_2(L_{per_1} + L_{per_2}) \\ &+ \lambda_3 L_{lan} + \lambda_4 L_{tv} + \lambda_5 L_{reg} + \lambda_6 L_{con} \\ &+ \lambda_7(L_{adv\_mid} + L_{adv\_high}),\end{aligned} \tag{1}$$

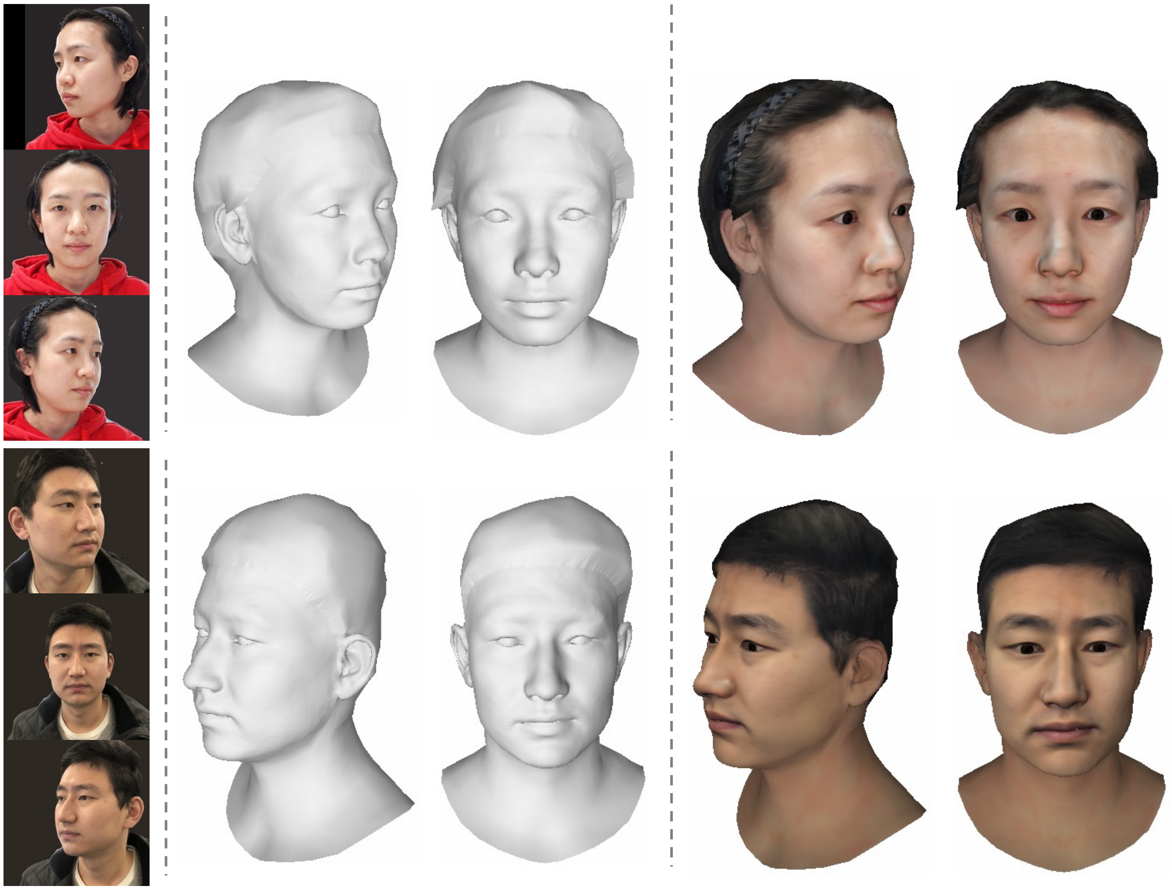
$$\mathcal{L}_{super} = L_{self} + \lambda_8(L_{mid} + L_{high}), \tag{2}$$

where $\lambda_1 = 1.9, \lambda_2 = 0.2, \lambda_3 = 1.6e-4, \lambda_4 = 5e3, \lambda_5 = 10, \lambda_6 = 20, \lambda_7 = 0.2$ and $\lambda_8 = 1$ as default. We alternately train the network with L_{self} for one iteration and with L_{super} for one iteration.

4. Head Reconstruction

Due to the lack of completeness, the application scenarios of face reconstruction are often limited. Therefore, we extend our method and make a small modification to MV-HRN to achieve high-quality head reconstruction. Firstly, we combine BFM with FLAME and generate a new head 3DMM model. Fig. 5 shows the pipeline. Given a face model from BFM database, we firstly use a template FLAME model and apply the flame-fitting [11] algorithm to fit the face model. Then through a series of cropping and merging operations, we can get the complete head model that combines BFM and FLAME. By applying the process above to the mean model, 80 identity blendshapes and 64 expression blendshapes of BFM, we can obtain a new head 3DMM, which shares the same group of coefficients with BFM. We continue to use the albedo basis of BFM for the face area of the new head models. Since we will only calculate the photometric loss for the face area, the albedo of the rest area of the head model is set to a fixed value. In addition, we unwrap the new head model and recalculate a new set of UV coordinates.

To adapt the new head model, we modify the MV-HRN by splitting the deformation map into the face deformation



(a) 3-view images (b) Predicted mesh (c) Textured mesh

Figure 7. Some head reconstruction results of our method on selfie data.

map and hair deformation map, where the former is used to deform the face region and the latter is used to deform the hair region. Besides, we replaced the face mask in contour-aware loss with the head mask, which is predicted by a pre-trained head segmentation network [12]. By using the face and hair deformation map to fit the head mesh to the head masks, we can get a head model that is well aligned to the input multi-view head images. Note that since there are no 3D priors to guide the deformation of the hair region, we apply a larger weight of L_{tv} for the hair deformation map to ensure the smoothness of the hair region. Fig. 6 shows a simplified process of reconstructing the head mesh using MV-HRN. For the texture, we firstly acquire the coarse texture maps from each view by employing the differentiable rendering [10] mentioned in Sec. 3.2 (main paper). Then we blend the multi-view textures and a template head texture following [7] to obtain the complete head texture map. Combining the predicted head mesh and the head texture map, we achieve high-quality head reconstruction from sparse-view images. Fig. 7 shows some head reconstruction results of our method on selfie data.

Due to the deficiency of prior information on hair regions and the limitation of mesh density, our head reconstruction is currently only suitable for handling portraits with simple hairstyles.

Method	Cooperative		Indoor		Outdoor	
	Mean	Std.	Mean	Std.	Mean	Std.
Tran et al. [14]	1.93	0.27	2.02	0.25	1.86	0.23
Booth et al. [2]	1.82	0.29	1.85	0.22	1.63	0.16
Genova et al. [6]	1.50	0.13	1.50	0.11	1.48	0.11
GANFit [5]	0.95	**0.107**	0.94	0.106	0.94	0.106
Ours	**0.86**	0.108	**0.87**	**0.105**	**0.83**	**0.104**

Table 1. The multi-view quantitative results on the MICC Dataset using point-to-plane distance (mm).

5. More Results

5.1. Comparisons on the MICC dataset

We follow GANFit and test our method on the MICC dataset [1]. Table 1 gives the results.

5.2. More Visualization Results

We provide qualitative comparison of single-view face reconstruction results with more methods (FDS [3] and LAP [15]) in Fig. 8, Fig. 9, and Fig. 10. Our approach consistently outperforms other methods on FFHQ, REALY and Facescape datasets with high-fidelity and fine details.

6. Limitations and Future Work

Limitations. We summarize two limitations of our method. On one hand, the generated facial details of our method are static and cannot vary with expression. One possible way is to collect multiple expressions of the same person, and then use HRN to obtain mid- and high-frequency details from each expression and build a set of mid- and high-frequency detail blendshapes. Finally, we are enabled to use blendshape coefficients to generate dynamic facial details.

On the other hand, considering the pixel-wise learning strategy and the powerful representation ability, our proposed method cannot handle severe occlusions well. Fig. 11 presents some visual results of the proposed HRN on occluded face images from FFHQ. Our method exhibits robustness to some images with slightly occluded faces (first two rows) but produces inaccurate results for heavily occluded faces (last three rows).

Future Work. Beyond addressing the limitations discussed above, we will further extend our method to achieve accurate, high-fidelity and animatable head avatar generation from in-the-wild images for future work, conquering some challenging problems (such as modeling complex hairstyles).

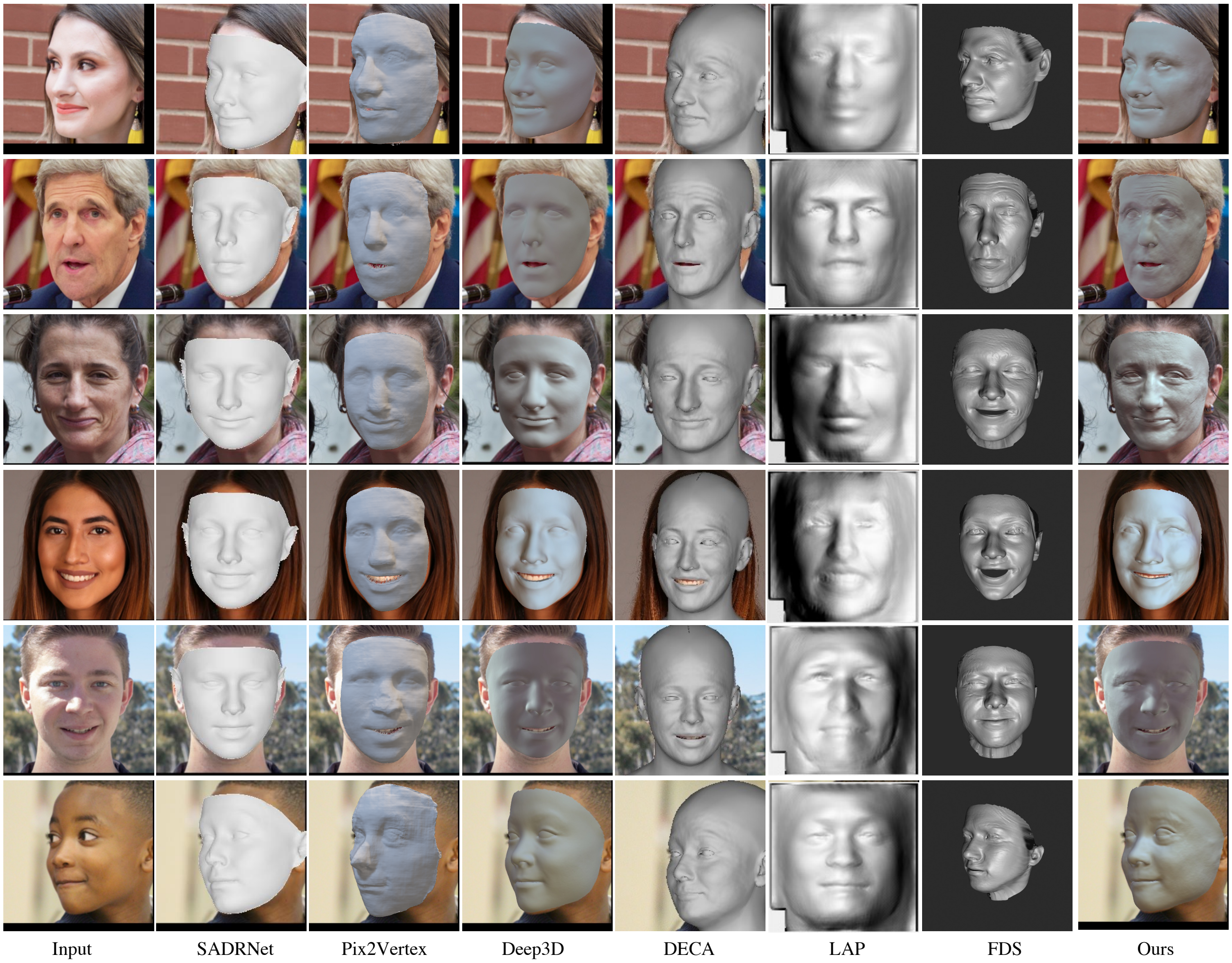


Figure 8. More visual comparisons on FFHQ dataset.

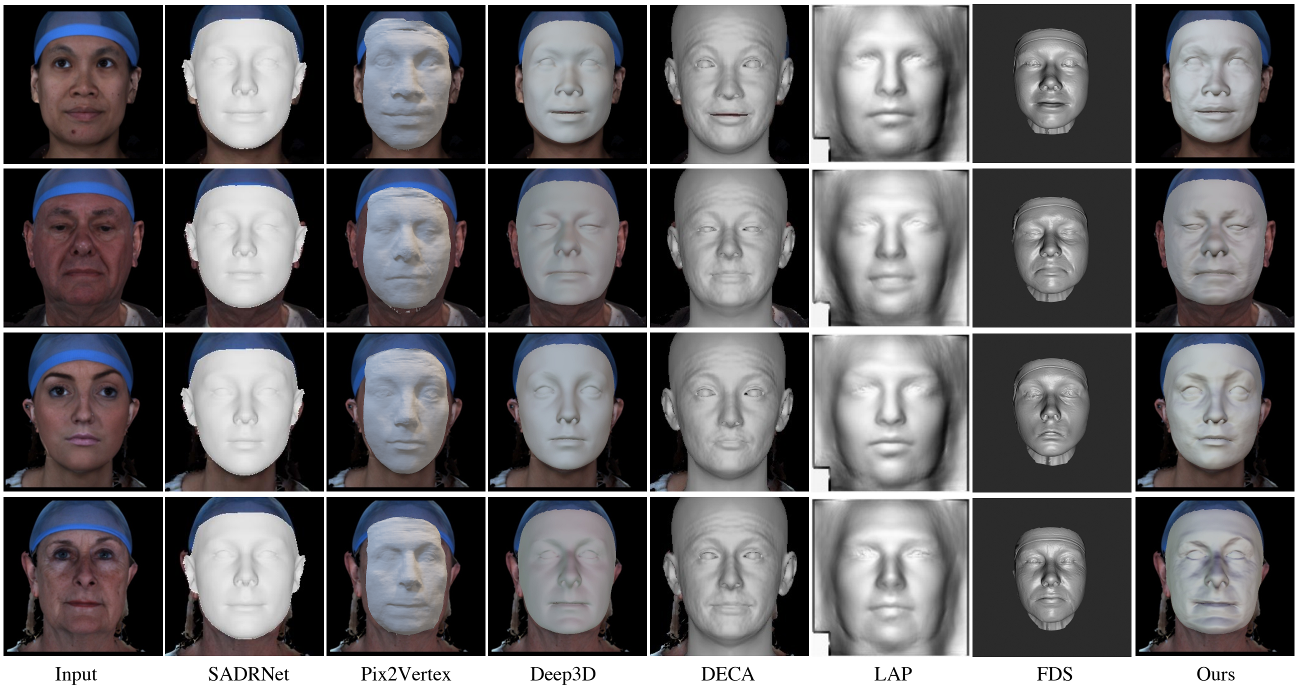


Figure 9. More visual comparisons on REALY dataset.

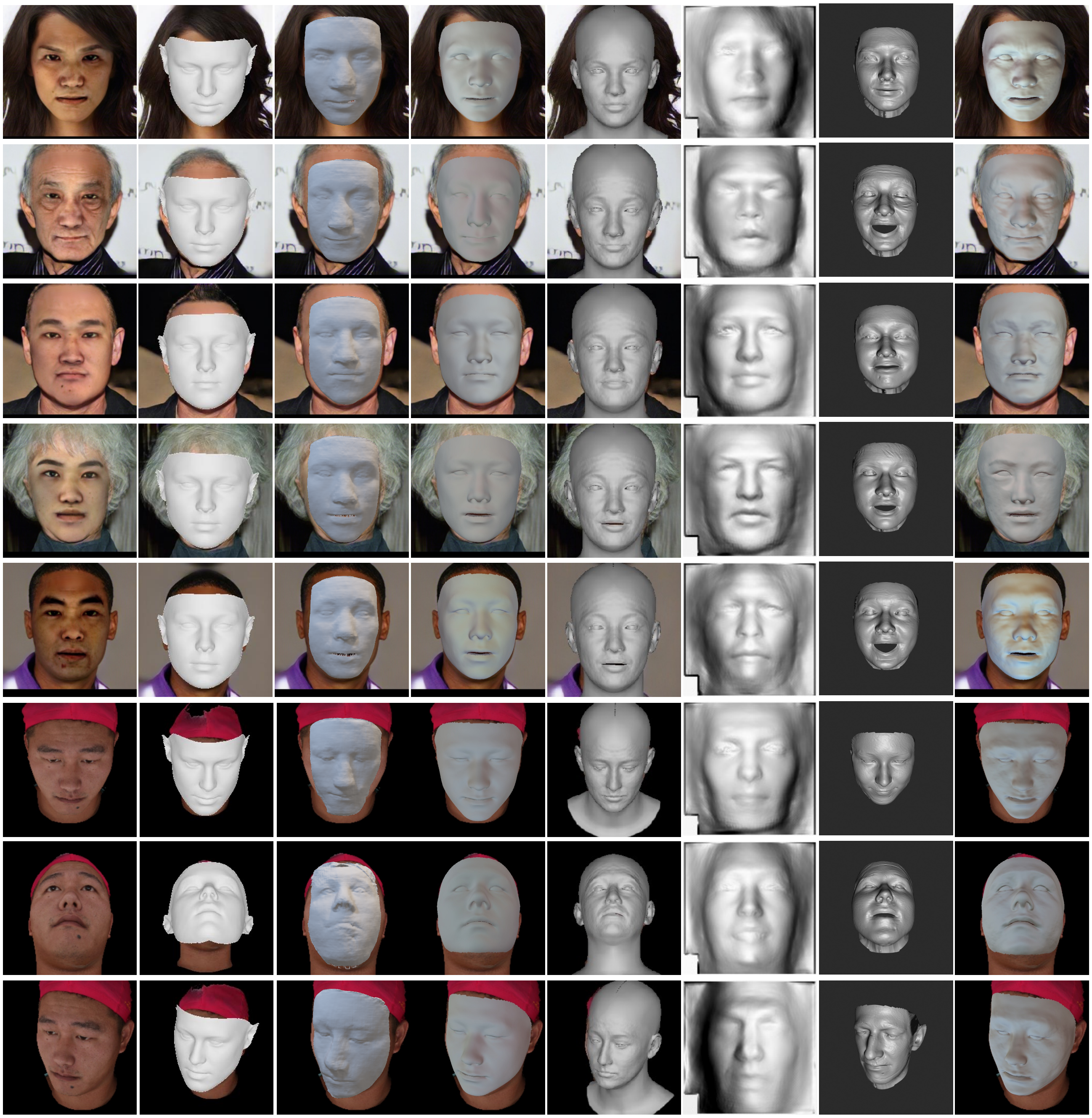
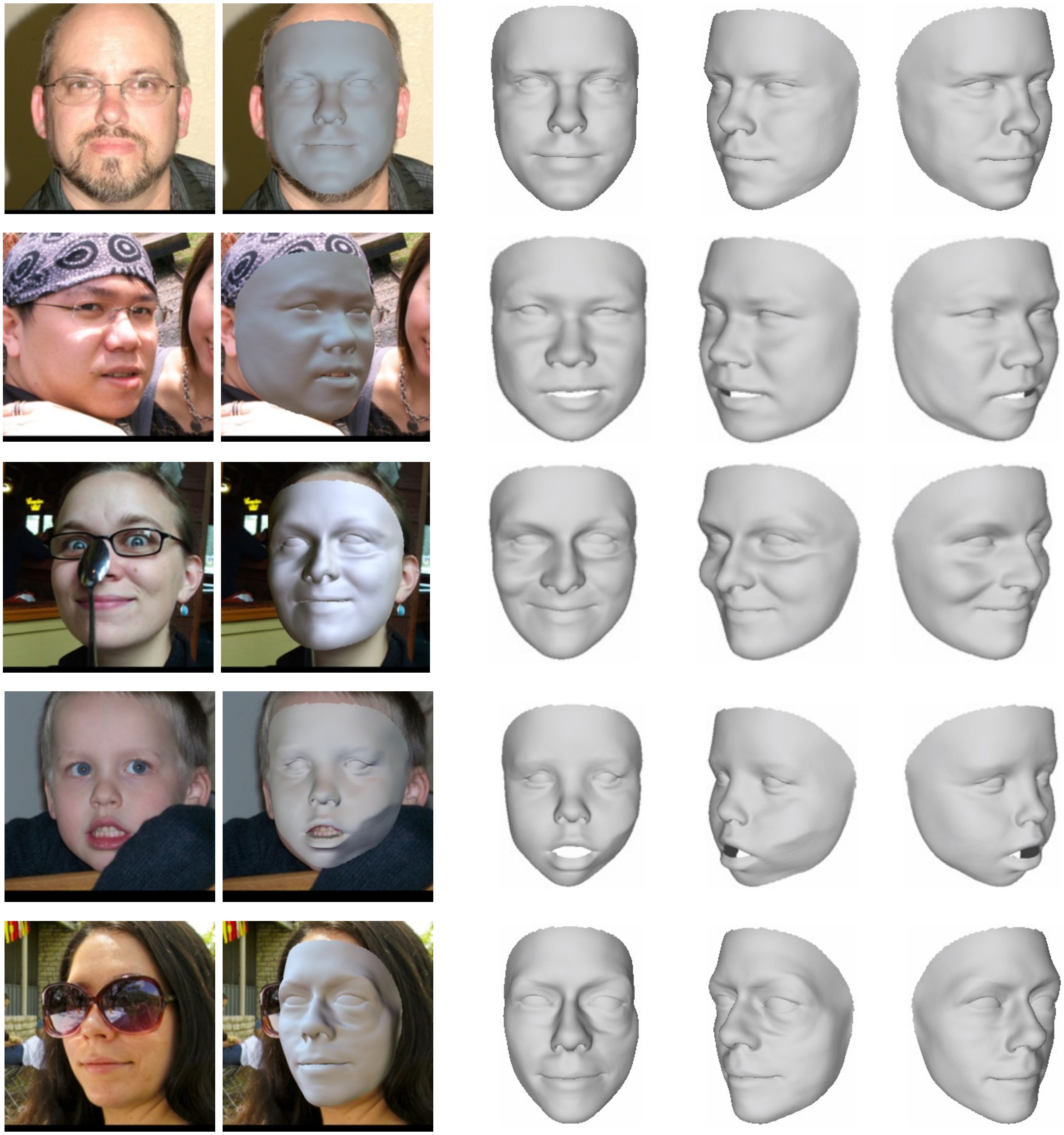


Figure 10. More visual comparisons on FaceScape dataset.



(a) Input (b) Ours (c) Predicted mesh

Figure 11. Visual results of our method on some occluded face images from FFHQ.

References

[1] Andrew D Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 79–80, 2011. 4

[2] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models" in-the-wild". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 48–57, 2017. 4

[3] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9429–9439, 2019. 4

[4] 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. 3

[5] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[6] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 4

[7] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):1–14, 2015. 4

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[10] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 4

[11] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 3

[12] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8572, 2020. 4

[13] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. 2

[14] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. 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*, pages 5163–5172, 2017. 4

[15] Zhenyu Zhang, Yanhao Ge, Renwang Chen, Ying Tai, Yan Yan, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning to aggregate and personalize 3d face from in-the-wild photo collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14214–14224, 2021. 4

[16] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 2