Supplemental Material for "Self-Supervised Monocular 3D Face Reconstruction by Occlusion-Aware Multi-view Geometry Consistency"

Jiaxiang Shang[1][0000−0001−7161−9765], Tianwei Shen[1][0000−0002−3290−2258], Shiwei li[1][0000−0003−0712−0059], Lei Zhou[1][0000−0003−4988−5084], Mingmin Zhen[1][0000−0002−8180−1023], Tian Fang[2][0000−0002−5871−3455], and Long Quan[1][00000001−8148−1771]

[1] Hong Kong University of Science and Technology
{jshang,tshenaa,lzhouai,mzhen,quan}@cse.ust.hk
[2] Everest Innovation Technology
{sli,fangtian}@altizure.com

1 Overview

This supplementary document provides detailed evaluation results that are supplementary to the main paper. We propose a self-supervised Multi-view Geometry Consistency based 3D Face Reconstruction framework (MGCNet), which helps mitigate the monocular face pose and depth ambiguity. Firstly, we propose the detailed data pre-process pipeline in Section 2, then we introduce the quantitative evaluation datasets in Section 3. Secondly, we introduce the morphable model and highlight that our MGCNet is a general framework in Section 4. Thirdly, we evaluate the quantitative result by render error between the input image and rendered image in Section 5.1 and we show the qualitative ablation study in Section 5.2. Furthermore, we show further comparison with Tewari19 [13] under geometry, texture and lighting in Section 5.3, then we conduct further comparison with some methods on the in the wild images in Section 5.4. Finally, we demonstrate the qualitative comparisons against other methods on MICC Florence dataset [1] in Section 5.5 and we also demonstrate some results from AFLW20003D [23] in Section 5.6, which further certify our MGCNet performs accurate result on face alignment task. Then we show the qualitative result on BU-3DFE dataset [19, 21] in Section 5.7.

2 Data Preprocess

The images are automatically annotated by the 2D landmark detection method in [3] and the face detection method in [20]. We filter the face pose, face attribution, low-resolution images, and blurred images and obtain ∼390K face images from the above four datasets as our training set. The images are scaled to a resolution of 224 × 224.

The multi-view images of the training dataset are captured with a consistent lighting condition across views. Theoretically, the photometric consistency will be violated if the lighting across views is dramatically different. Our MGCNet shares the same property with multi-view stereo methods that require overlap across views, this also the reason that we only use $N = 3$ views in practice.

3 Quantitative Evaluation Dataset

AFLW20003D is constructed to evaluate face alignment on challenging in the wild images. This database contains the first 2000 images from AFLW [7] with landmarks annotations. We use this database to evaluate the performance of our method on face alignment tasks [23].
MICC Florence is a 3D face dataset that contains 53 faces with their ground truth High-resolution 3D scans of human faces are acquired from a structured-light scanning system from each subject with several video sequences of varying resolution, conditions and zoom level [1].
FRGC v2.0 includes 4007 scans of 466 individuals acquired with the frontal view from the shoulder level, with very tiny pose variations. About 60% of the faces have neutral expression, while the others show spontaneous expressions of disgust, happiness, sadness, and surprise [9]. Scans are given as matrices of 3D points of size $[480, 640]$, with a binary mask indicating the valid points of the face (about 40 K on average).
BU-3DFE BU-3DFE database includes 100 subjects with 2500 facial expression models. The database presently contains 100 subjects (56% female, 44% male), ranging age from 18 years to 70 years old, with a variety of ethnic/racial ancestries, including White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino. Each subject performed seven expressions in front of the 3D face scanner. With the exception of the neutral expression, each of the six prototypic expressions (happiness, disgust, fear, angry, surprise and sadness) includes four levels of intensity [19, 21].

4 3D Morphable Model

Blanz and Vetter [2] introduce the 3D morphable model (3DMM). 3DMM benefits the 3D face reconstruction by constraining the solution space, thereby simplifying the problem. In this paper, our goal is to estimate 3DMM parameters from a single photograph.

We conduct our experiments with 3DMM model since it is still a general method that widely used by single-image based latest methods (as in works of [4, 6, 15, 18, 22, 23]), and we have proved the superiority of our method over theirs under a fair comparison, as shown qualitatively in the main paper as well as our quantitative result.

As we have clarified in the main paper, our method is focused on improving single-view reconstruction quality via multi-view consistency, our proposed

framework is general and is not limited to any specific face model. We believe other face models with better representation ability [11, 13, 14, 16, 17] can easily plug into our proposed MGCNet.

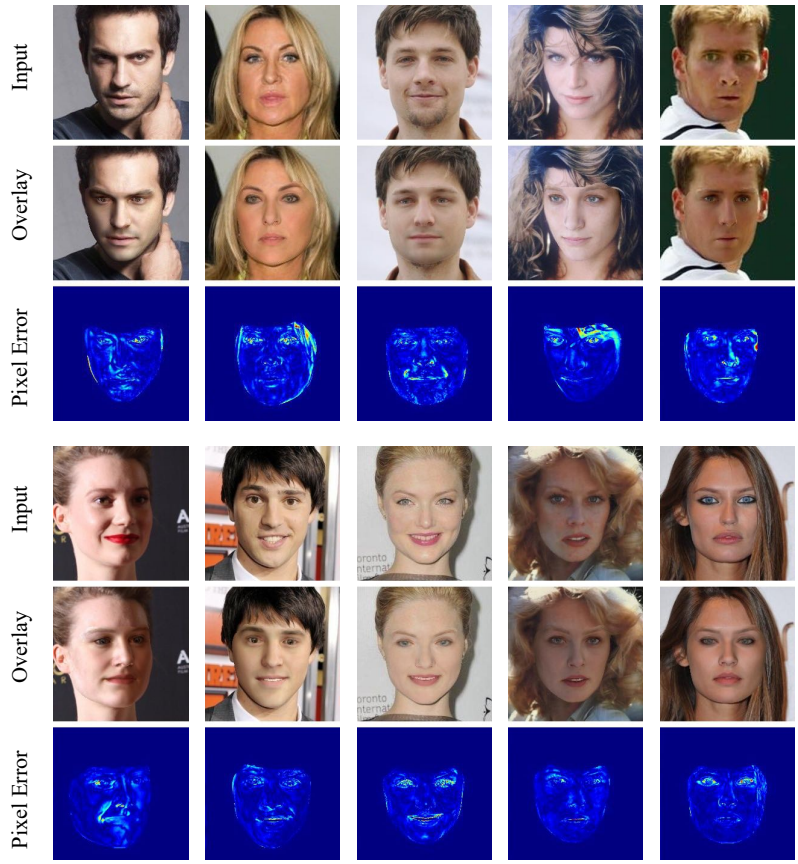

Fig. 1. Quantitative evaluation of photometric error on the CelebA [8] dataset. The error map range is $[0, 1]$

5 Further Evaluation Result

5.1 CelebA Dataset

We evaluate the photometric error of our approach by heat maps on the CelebA dataset [8] as shown in Figure 1 that these images are only used for testing and visualization. We achieve low pixel error, which benefits from using multi-view geometry consistency. This also demonstrates better reconstruction capabilities of our MGCNet to in-the-wild images.

5.2 Alation Study

To evaluate the effects of multi-view geometry consistency losses on the quality of the reconstructed meshes. We conduct ablation studies on the MICC Florence 3D Face dataset [1], as shown in Figure 2. We calculate the point-to-plane root mean squared error, and normalize the error to a heatmap. This heatmap indicate that the major improvements regions are jaw, nose and cheekbones region in frontal case, and eye contour, nose, cheekbones regions for the large-pose case. Face geometry (especially in large pose cases), as well as better 3D pose estimation results, are the major improvements bring by our method, thanks to our multi-view geometry constraints that explicitly regularizes the geometry across different views.

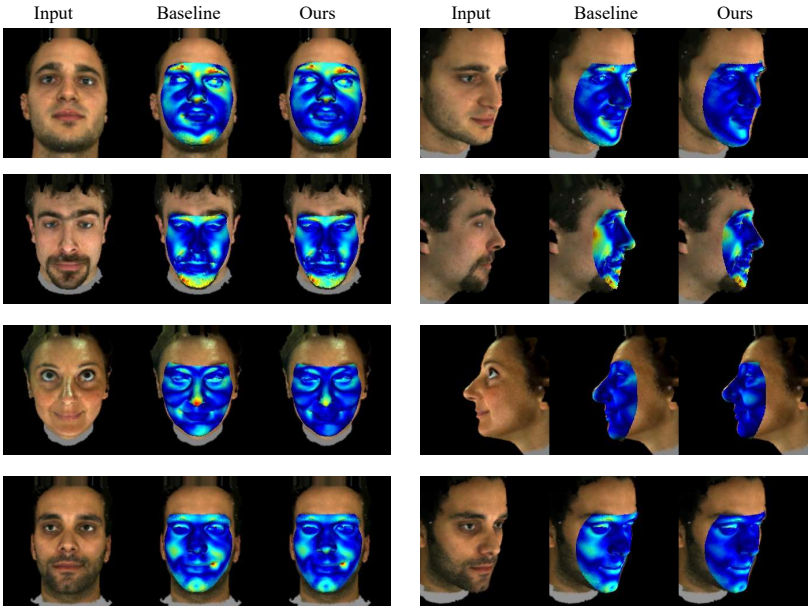

Fig. 2. Quantitative evaluation of point-to-plane root mean squared error as the error map format on the MICC Florence 3D Face dataset [1]. The error map range is $[0, 8.29]$.

5.3 Texture, illumination shadings

We compare our MGCNet with Tewari19 [13] as Figure 3. The result of Tewari19 [13] results of geometry are visually more detail since they use a face representation more complicated than 3DMM used by our method. However, it is hard to say Tewari19 [13] has better geometry results since our method does have a better quantitative result shown in the main paper. The texture model used in

Tewari19 [13] is also different from 3DMM. Despite that, better geometry generated by our method leads to better texture via the render loss used, which can also support the validity of our MGCNet. As our method is focused on improving the reconstruction quality via multi-view consistency. Our MGCNet is a general system that is not limited to any specific face model.

5.4 In the wild data

Secondly, we visualize our result under geometry overlay situation compared with Richardson *et al.* [10], Sela *et al.* [12], Tewari17 *et al.* [15], Tewari19 *et al.* [13] and RingNet *et al.* [11], and we notice that our approach performs better than methods [10–13, 15] as shown in Figure 5.

We also show some detail intermediate result about *Figure 5 in the main paper* in Figure 7.

5.5 MICC Florence Dataset

Firstly, we compare our MGCNet with Zhu *et al.* [23] (3DDFA), Sanyal *et al.* [11] (RingNet), Feng *et al.* [5] (PRN), and Deng *et al.* [4]. For each sample in MICC Florence, we pick both front face images and large face pose images as test data. We show the geometry overlay of the reconstruction result, which we achieve more accurate results than the most methods, and we get better results than Deng *et al.* [4] in the large pose case as Figure 4.

5.6 AFLW20003D Dataset

AFLW20003D is constructed by [23] to evaluate face alignment. Since the images are captured in the wild and show large variations in pose and appearance, which is a challenging 3D face alignment dataset. We use this database to evaluate the performance of our method on face alignment tasks.

As a supplementary to the quantitative evaluation in the main paper, we first demonstrate some results even better than the ground truth from AFLW20003D [23] in Figure 6. Besides, we also show our result that performs accurate face alignment results, where red lines are predicted landmarks by our method, white lines are ground truth from [23].

Furthermore, we visual our MGCNet result from multiple viewpoints in Figure 8 on AFLW20003D [23], which shows that we get vivid reconstruction results.

5.7 BU-3DFE Dataset

We present more qualitative reconstruction results of our MGCNet on BU-3DFE dataset [19,21]. In Figure 9, we show six samples with various expressions, the reconstructed 3D face showed with face pose, texture, geometry and illumination. This quantitative evaluation of our geometry reconstruction on the BU-3DFE dataset [19,21] shows that our MGCNet can even handle different expression situation. For the prediction of albedo and lighting, while the texture quality would

be benefited from better geometry implicitly via the render loss, the ambiguity in illumination and face albedo is an intrinsic issue due to the problems nature, and our SH lighting is RGB-channel, limit the SH lighting to one channel will help.

References

1. Bagdanov, A.D., Del Bimbo, A., Masi, I.: The florence 2d/3d hybrid face dataset. In: Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding. p. 7980. J-HGBU 11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2072572.2072597, http://doi.acm.org/10.1145/2072572.2072597
2. Blanz, V., Vetter, T., et al.: A morphable model for the synthesis of 3d faces. In: Siggraph. vol. 99, pp. 187–194 (1999)
3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1021–1030 (2017)
4. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
5. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 534–551 (2018)
6. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8377–8386 (2018)
7. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops). pp. 2144–2151. IEEE (2011)
8. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
9. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 947–954. IEEE (2005)
10. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1259–1268 (2017)
11. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7763–7772 (2019)
12. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1576–1585 (2017)

13. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Fml: face model learning from videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10812–10822 (2019)
14. Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2549–2559 (2018)
15. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1274–1283 (2017)
16. Tran, L., Liu, F., Liu, X.: Towards high-fidelity nonlinear 3d face morphable model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1126–1135 (2019)
17. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7346–7355 (2018)
18. Tuan Tran, A., Hassner, T., Masi, I., Medioni, G.: 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. pp. 5163–5172 (2017)
19. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06). pp. 211–216. IEEE (2006)
20. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: S3fd: Single shot scale-invariant face detector. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 192–201 (2017)
21. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P.: A high-resolution spontaneous 3d dynamic facial expression database. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 1–6. IEEE (2013)
22. Zhou, Y., Deng, J., Kotsia, I., Zafeiriou, S.: Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1097–1106 (2019)
23. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016)

Fig. 3. Quantitative evaluation compare with Tewari19 [13].

Fig. 4. Comparison with Zhu *et al.* [23] (3DDFA), Sanyal *et al.* [11] (RingNet), Feng *et al.* [5] (PRN), and Deng *et al.* [4].

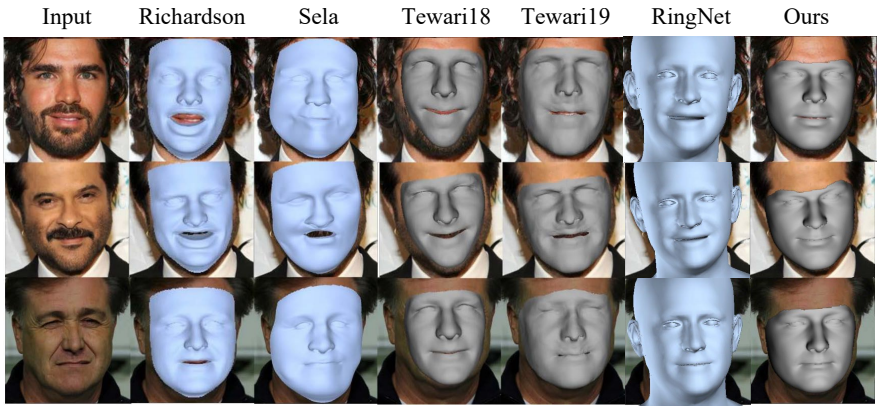

Fig. 5. Comparison with Richardson *et al.* [10], Sela *et al.* [12], Tewari17 *et al.* [15], Tewari19 *et al.* [13] and RingNet *et al.* [11]. Our MGCNet trained by multi-view consistency loss outperforms these state-of-the-art methods in face reconstruction geometry

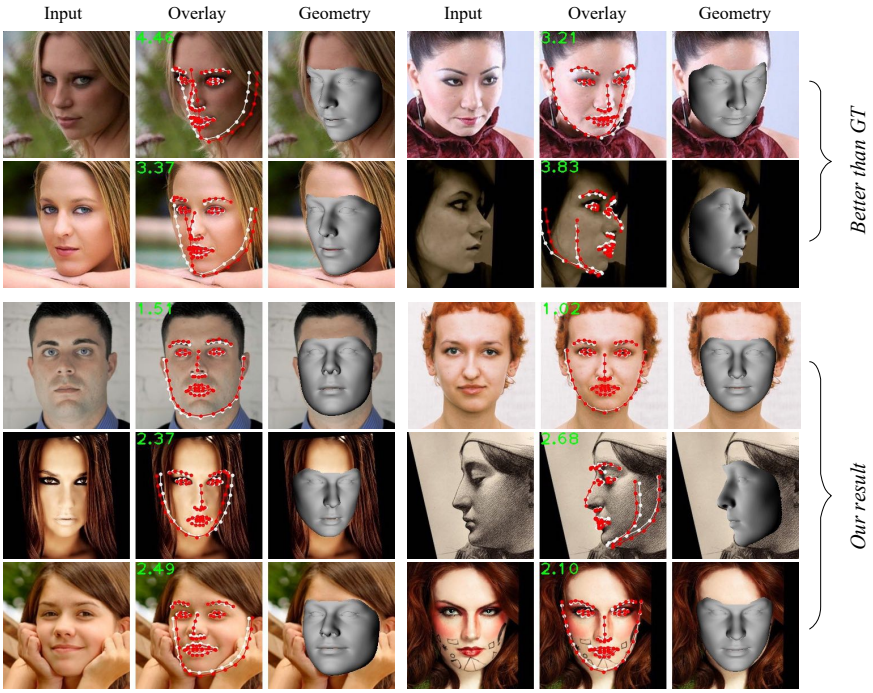

Fig. 6. Examples from AFLW20003D dataset [23] show that our predictions are more accurate than ground truth in some cases.

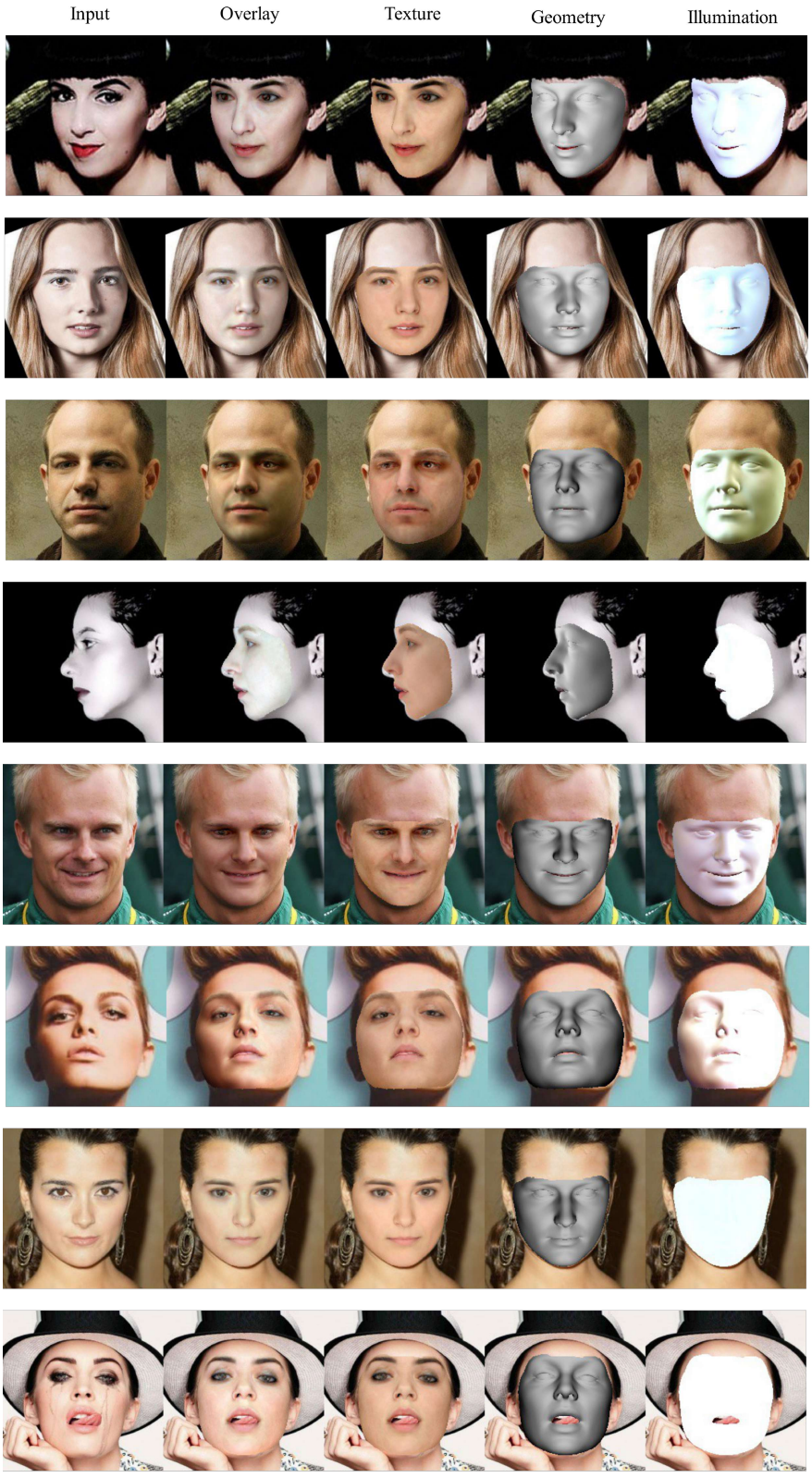

Fig. 7. Face reconstruction results under texture, geometry and illumination of our method on AFLW20003D [23] and CelebA [8].

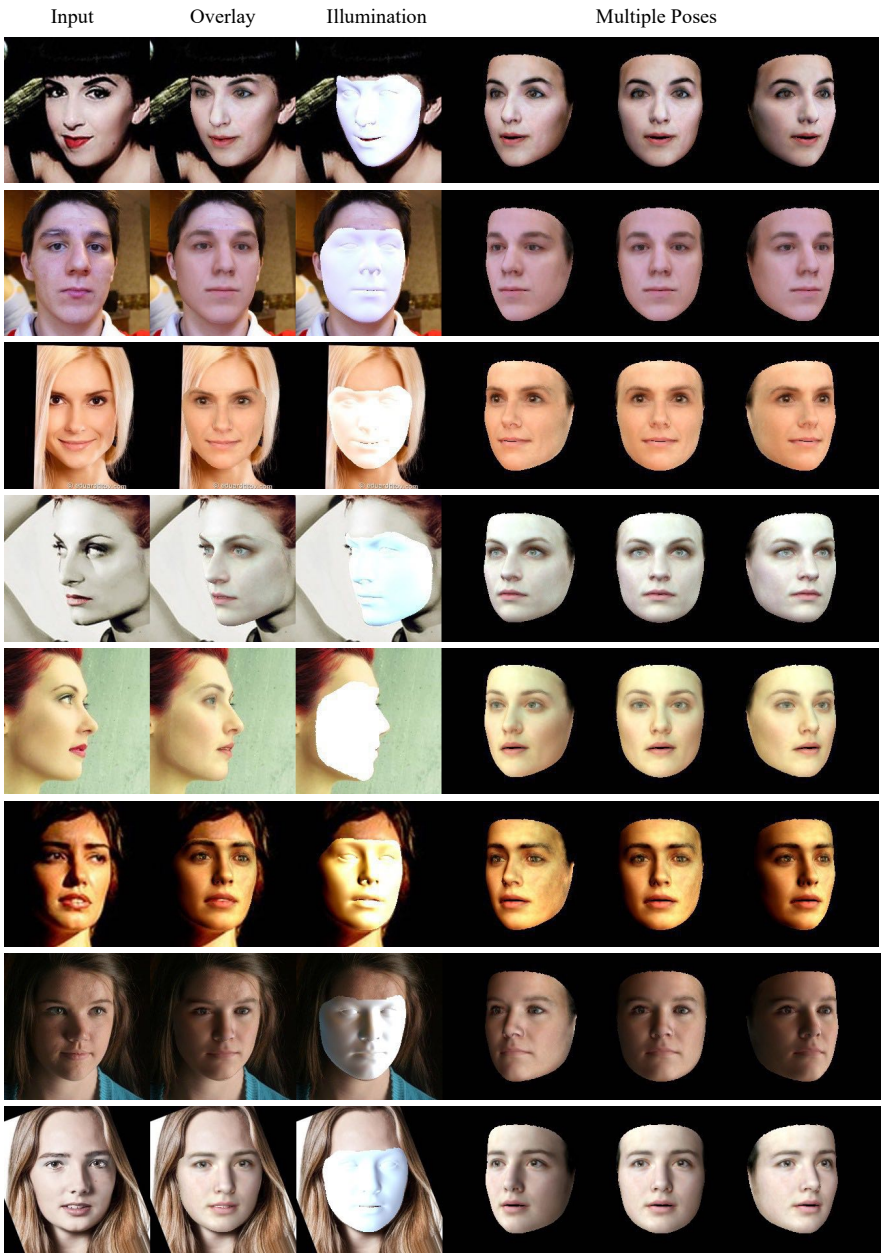

Fig. 8. Face reconstruction results of our method on AFLW20003D [23]

Fig. 9. Our accurate result in face pose, texture, geometry and illumination on BU-3DFE dataset [19, 21]