Supplemental Material for FFHQ-UV: Normalized Facial UV-Texture Dataset for 3D Face Reconstruction

Haoran Bai[1]* Di Kang[2] Haoxian Zhang[2] Jinshan Pan[1]† Linchao Bao[2]
[1]Nanjing University of Science and Technology [2]Tencent AI Lab

In this supplementary material, we first provide more analysis of the proposed data creation pipeline (Sec. 1). Then, we show more analysis of the proposed 3D face reconstruction algorithm, including quantitative evaluations, visual comparisons with state-of-the-art methods, and renderings under realistic conditions (Sec. 2). Finally, we show more examples of the proposed FFHQ-UV dataset rendered under different realistic lighting conditions (Sec. 3).

1. More analysis of the data creation

As stated in Sec. 3.2 of the manuscript, we have analyzed the effectiveness of the three major steps (i.e., StyleGAN-based image editing, UV-texture extraction, and UV-texture correction & completion) of our dataset creation pipeline. In this supplemental material, we further provide more detailed analysis.

1.1. Analysis on facial image editing

To obtain normalized faces, we first apply StyleFlow [1] to normalize the lighting, eyeglasses, hair, and head pose attributes of the input faces. Then, we edit the facial expression attribute by walking the latent code along the found direction of editing facial expression. One may wonder why not directly use StyleFlow to normalize the facial expression attribute. To answer this question, we compare the method which uses StyleFlow to edit facial expression in Fig. 1. The results show that using StyleFlow cannot normalize the expression attribute well, resulting in texture UV-maps containing unwanted expression information. In contrast, our method is able to generate neutral faces and produce high-quality texture UV maps.

Furthermore, in Fig. 3, we show some intermediate results of the proposed StyleGAN-based facial image editing.

1.2. Analysis on UV-texture completion

After detecting the artifact masks, we first use Poisson editing [9] to correct the artifact regions and then use Laplacian pyramid blending [2] to handle the remaining non-face regions (e.g., ear, neck, hair, etc.). One may wonder why not directly use Poisson editing or Laplacian pyramid blending to process all regions. To answer this question, in Fig. 2, we show the comparisons on the results produced by only using Laplacian pyramid blending, only using Poisson editing, and using the proposed method. The results show that Laplacian pyramid blending cannot handle the facial regions (e.g., eyes, nostrils) well, because color matching operation usually fails in these regions. Although Poisson editing can generate decent results, the huge time consumption is not suitable for creating large-scale datasets. In contrast, the proposed method can achieve similar results to Poisson editing in a short time.

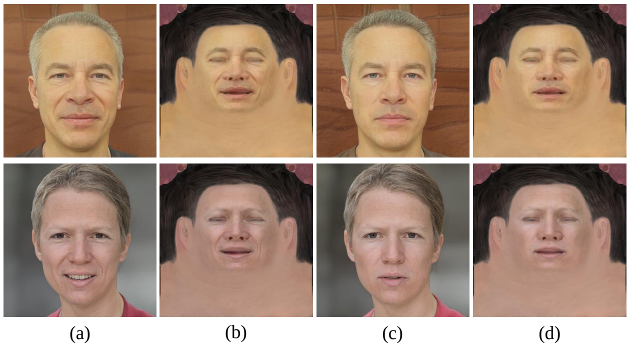

Figure 1. Comparisons on the results of editing expression attribute using StyleFlow [1] and our method. Note that using StyleFlow cannot normalize the expression attribute well (a), resulting in texture UV-maps containing unwanted expression information (b). In contrast, our method is able to generate neutral faces (c) and produce high-quality texture UV maps (d).

1.3. More visual comparisons

In Sec. 3.2.1 of the manuscript, we have shown an example of the UV-maps obtained by different baseline methods. In this supplemental material, we further provide more visual comparison in Fig. 4, which demonstrate the effectiveness of the three major steps of our dataset creation pipeline.

1.4. Analysis on data diversity

In Table 3 of the manuscript, we have shown the identity similarity computed between each image in FFHQ-Norm

*Work done during an internship at Tencent AI Lab.
†Corresponding author.

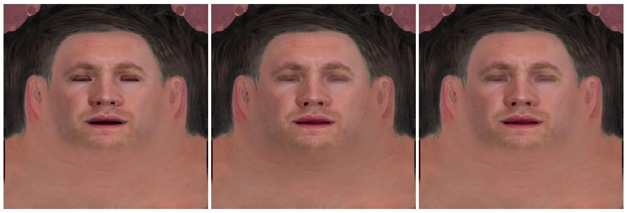
Lap. pyramid blending (Time: 0.93s)
Poisson editing (Time: 282.21s)
Ours (Time: 1.53s)

Figure 2. Comparisons on the results produced by only using Laplacian pyramid blending, only using Poisson editing, and using our method. Note that Laplacian pyramid blending cannot handle the facial regions (e.g., eyes, nostrils) well, and Poisson editing takes a huge amount of time to solve. In contrast, our method can achieve similar results to Poisson editing in a short time.

and the rendered image using the corresponding UV-map in FFHQ-UV with a random pose. In this supplemental material, we further compute the average identity similarity with images in the original FFHQ dataset in Tab. 1 (“w/ orig. FFHQ”), showing that our dataset creation pipeline preserves identity well. In addition, one may wonder whether using random pose renderings to compute identity features affects the computation of similarity scores. Thereby, we further provide the average identity similarity with frontal renderings in Tab. 1 (“w/ frontal pose”), where ours is still the best.

Table 1. Average ID similarity score.

Methods	negative samples	w/o multi-view	naive blending	Ours
w/ orig. FFHQ	0.0710	0.3723	0.4092	**0.4262**
w/ frontal pose	0.0594	0.8366	0.8481	**0.8520**

2. More analysis of 3D face reconstruction

2.1. Evaluation on Facescape dataset

As stated in Sec. 4.3 of the manuscript, we have evaluated the shape reconstruction accuracy on REALY benchmark [3]. In this supplemental material, we further evaluate the proposed 3D face reconstruction algorithm on the recent public Facescape dataset [10] with corresponding 3D scans and ground-truth texture UV-maps that can be used to compute quantitative metrics. We apply the first 100 subjects from Facescape, and randomly select one picture from multi-view faces as input to evaluate the shape and texture of the monocular reconstruction. For shape accuracy, we compute the average point-to-mesh distance between the reconstructed shapes and the ground-truth 3D scans. For texture evaluation, we use the interactive nonrigid registration tool to align the ground-truth texture UV-map in Facescape to our topology, and use the mean L1 error as the metric.

Tab. 2 shows the quantitative comparisons on different results directly predicted by Deep3D in stage 1 (denoted as “$\mathcal{N}_{enc}$”), generated using linear PCA texture basis instead of the GAN-based texture decoder in Stage 2 & 3 (“PCA tex. basis”), generated using a texture decoder trained on the UV-map dataset created without generating multi-view images (“w/o multi-view”), and generated using the texture decoder trained on our final FFHQ-UV dataset (“Ours”).

Table 2. Quantitative comparison of reconstructed shapes and textures on the FaceScape dataset [10].

Methods	shape (mm) ↓ (point-to-mesh dist.)	texture ↓ (L1 error)
$\mathcal{N}_{enc}$	1.537	0.1719
PCA tex. basis	1.524	0.1706
w/o multi-view	1.502	0.1433
Ours	**1.495**	**0.1425**

The results indicate that the proposed 3D face reconstruction algorithm, based on the GAN-based texture decoder trained with the proposed FFHQ-UV dataset, is able to improve the reconstruction accuracy in terms of both shape and texture.

2.2. Evaluation on illumination and facial ID

In Sec. 3.2 of the manuscript, we have evaluated the illumination and facial identity preservation of the data creation. In this supplemental material, we further evaluate these on the proposed 3D face reconstruction algorithm. We test them on REALY benchmark [3] in Tab. 3. We first show the illumination evaluation of the reconstructed UV-maps (see Tab. 3 “BS Error”), then compute the identity similarity between the input faces and the rendered faces to verify whether the reconstructed face can preserve the identity (see Tab. 3 “Similarity”). Our method outperforms the baseline variants.

Table 3. More evaluations on 3D face reconstruction.

Methods	$\mathcal{N}_{enc}$	PCA tex basis	w/o editing	w/o multi-view	Ours
BS Error	-	-	8.850	4.683	**4.322**
Similarity	0.7832	0.7946	0.7153	0.7980	**0.8102**
Exp. Acc.	73%	74%	69%	78%	**82%**

2.3. Evaluation on facial expression

In this section, we further evaluate facial expression preservation using the expression classifier. As the expressions of faces in REALY [3] are all neutral, we further collect 100 faces with different expressions from the websites. We calculate the consistency of the two predictions (i.e. the original input images and the rendered images using the reconstructed shape/texture) from the expression classifier. Results in Tab. 3 (“Exp. Acc.”) show that our reconstruction method can better preserve the expressions.

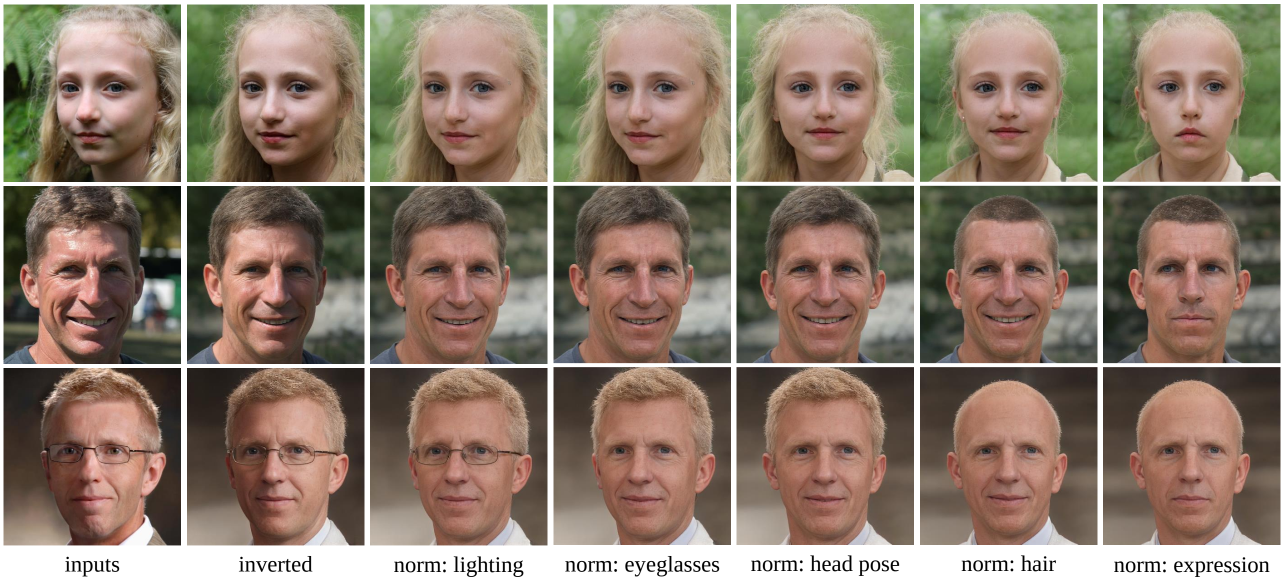

Figure 3. Intermediate results of StyleGAN-based facial image editing, where the lighting, eyeglasses, head pose, hair, and expression attributes of the input faces are normalized sequentially.

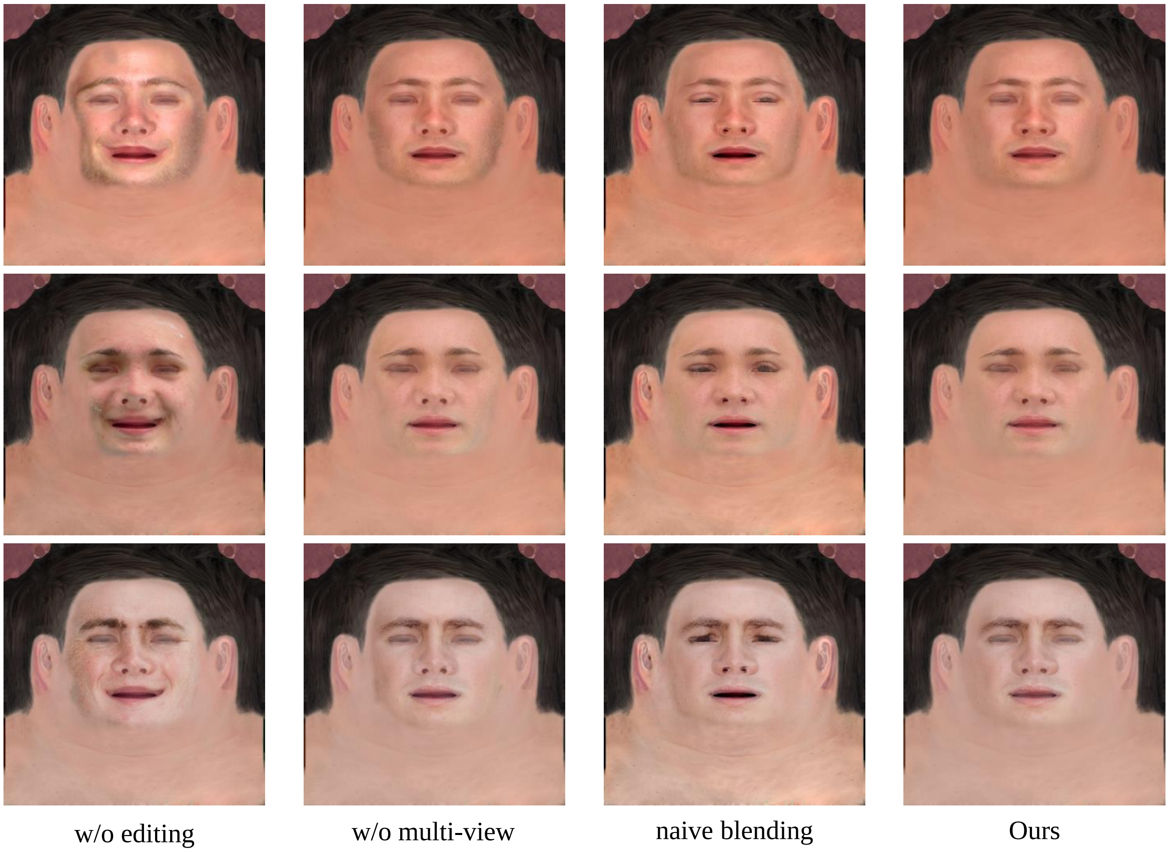

Figure 4. Extracted UV-maps by different baseline methods of our data creation pipeline, where our method is able to produce higher quality textures.

2.4. More visual results

In this section, we show more visual comparisons of the reconstructed faces produced by different methods, including GANFIT [5], AvatarMe [6], Normalized Avatar [6], HQ3D-ACN [7], StyleFaceUV [4], and our method. As shown in Fig. 5-6, meshes and texture UV-maps reconstructed by our method are superior to other results in terms of both fidelity and quality. Compared to Normalized Avatar [6], as shown in Fig. 7, our results better resemble the input faces, and our method is able to express more skin tones, thanks to the more powerful expressive texture decoder trained on our much larger dataset. In Fig. 8, we show the visual comparisons of the reconstruction results to recent approaches HQ3D-ACN [7] and StyleFaceUV [4], where our results better resemble the input faces.

In Fig. 9-11, we further show more examples of our reconstructed texture UV-maps, shapes, and renderings under different lighting conditions.

3. More examples of FFHQ-UV dataset

In this section, we provide more examples of the proposed FFHQ-UV dataset in Fig. 12-13. The produced texture UV-maps are with even illuminations, neutral expressions, and cleaned facial regions (e.g., no eyeglasses and hair). Thus, they are ready for rendering under different lighting conditions. In our visualizations, there realistic environment lighting conditions are demonstrated including a studio scene[*], a garden scene[†], and a chapel scene[‡]).

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 2021. 1

[2] Peter J Burt and Edward H Adelson. A multiresolution spline with application to image mosaics. *ACM Trans. Graph.*, 1983. 1

[3] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. REALY: Rethinking the evaluation of 3D face reconstruction. In *ECCV*, 2022. 2

[4] Wei-Chieh Chung, Jian-Kai Zhu, I-Chao Shen, Yu-Ting Wu, and Yung-Yu Chuang. Stylefaceuv: A 3d face uv map generator for view-consistent face image synthesis. pages 89–99, 2022. 4, 8

[5] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFit: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *CVPR*, 2019. 4, 5, 6

[6] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. AvatarMe: Realistically renderable 3D facial reconstruction in-the-wild. In *CVPR*, 2020. 4, 5, 6

[7] Zhiqian Lin, Jiangke Lin, Lincheng Li, Yi Yuan, and Zhengxia Zou. High-quality 3d face reconstruction with affine convolutional networks. In *ACM MM*, pages 2495–2503, 2022. 4, 8

[8] Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. Normalized avatar synthesis using StyleGAN and perceptual refinement. In *CVPR*, 2021. 7

[9] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 2003. 1

[10] 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 *CVPR*, pages 601–610, 2020. 2

[*] https://polyhaven.com/a/studio_small_09

[†] https://polyhaven.com/a/garden_nook

[‡] https://polyhaven.com/a/thatch_chapel

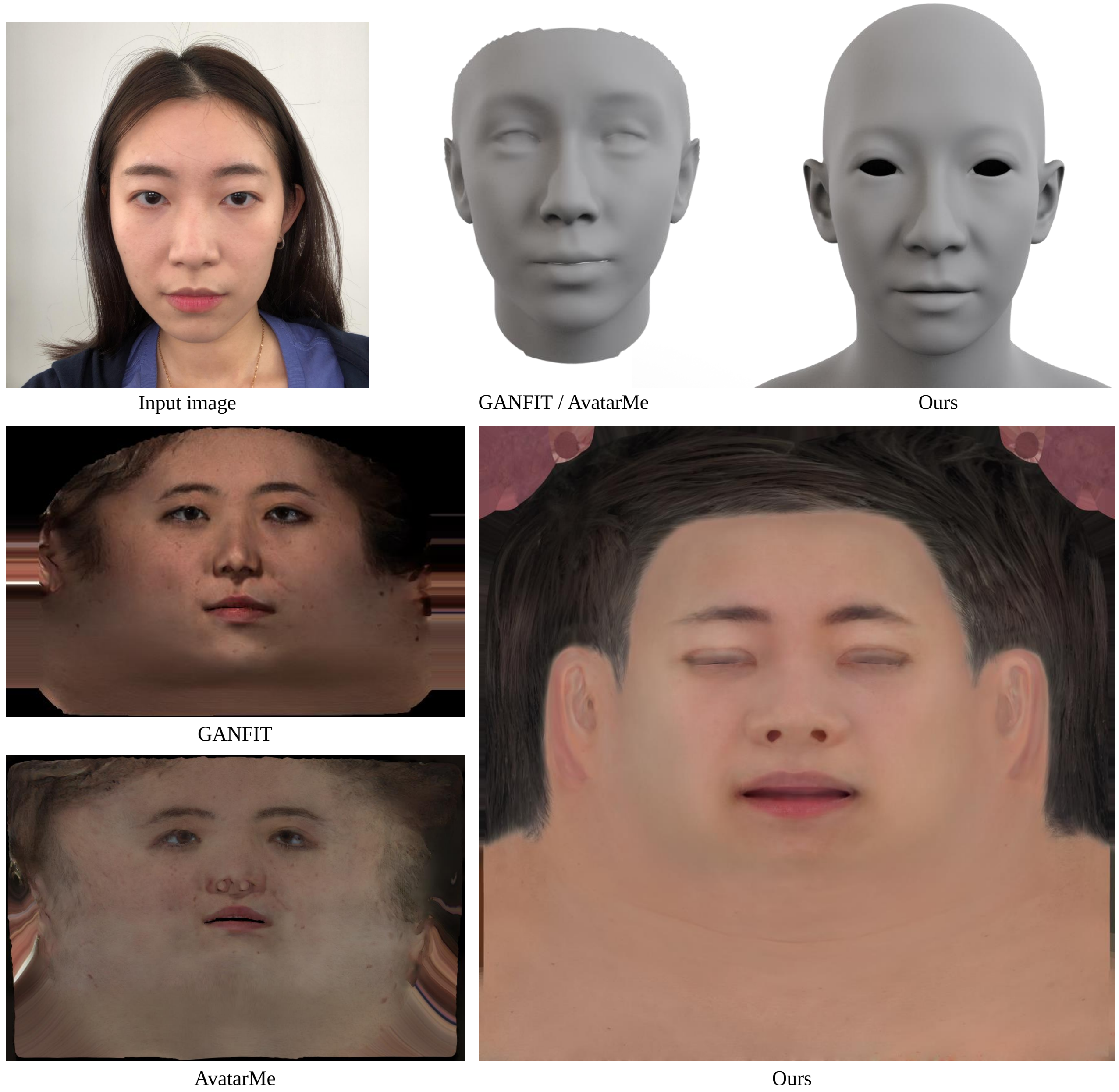

Figure 5. Visual comparisons of the reconstruction results to state-of-the-art approaches GANFIT [5] and AvatarMe [6], where our reconstructed shape is more faithful to the input face. Note that there are undesired shadows and uneven shadings in the UV-maps obtained by GANFIT and AvatarMe, while our UV-map is more evenly illuminated and of higher quality.

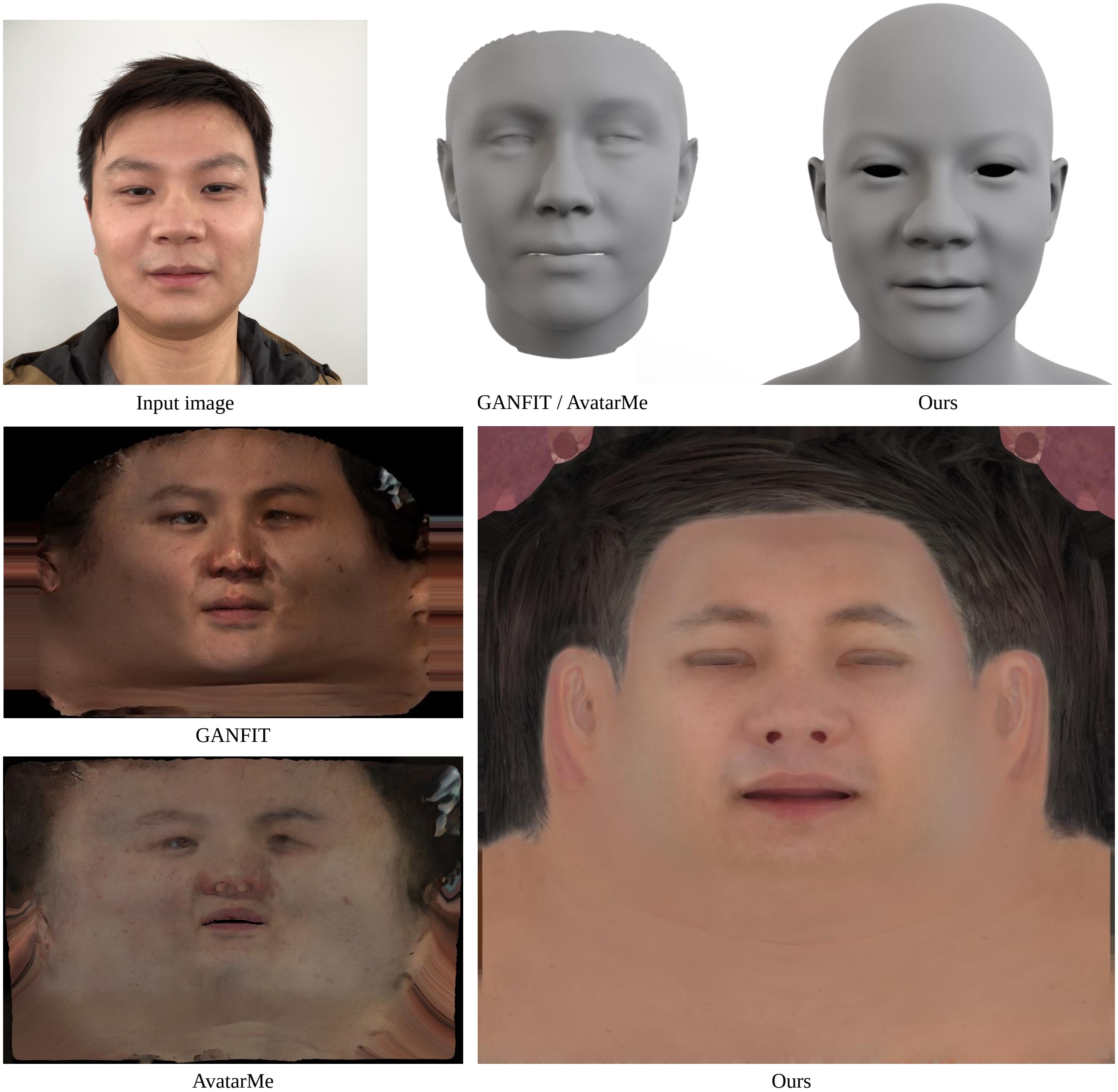

Figure 6. Visual comparisons of the reconstruction results to state-of-the-art approaches GANFIT [5] and AvatarMe [6], where our reconstructed shape is more faithful to the input face (e.g., nose region). Note that there are undesired shadows and uneven shadings in the UV-maps obtained by GANFIT and AvatarMe, while our UV-map is more evenly illuminated and of higher quality.

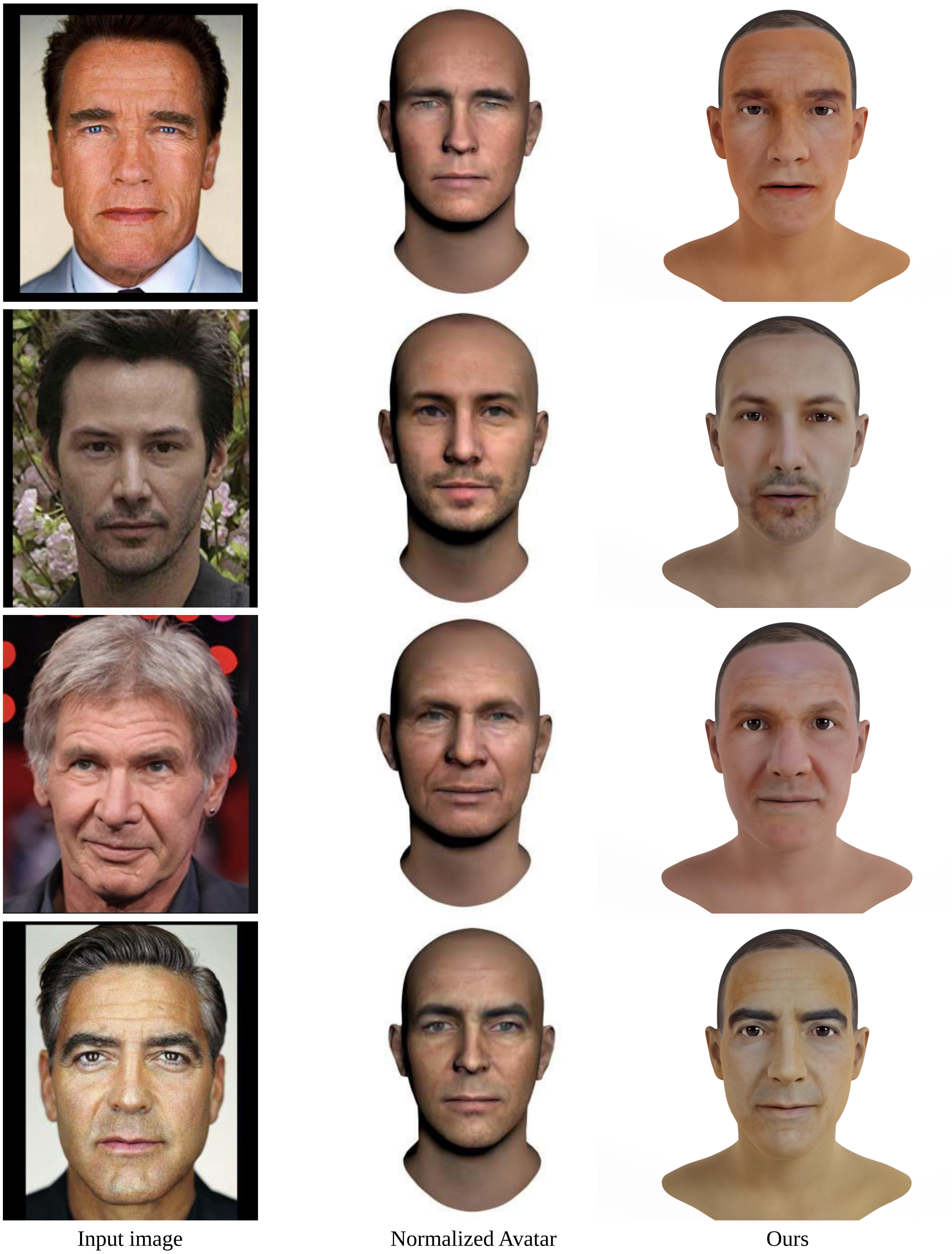

Figure 7. Visual comparisons of the reconstructions between Normalized Avatar [8] and ours, where our results better resemble the input faces. Note that our method is able to express more skin tones, thanks to the more powerful expressive texture decoder trained on our much larger dataset.

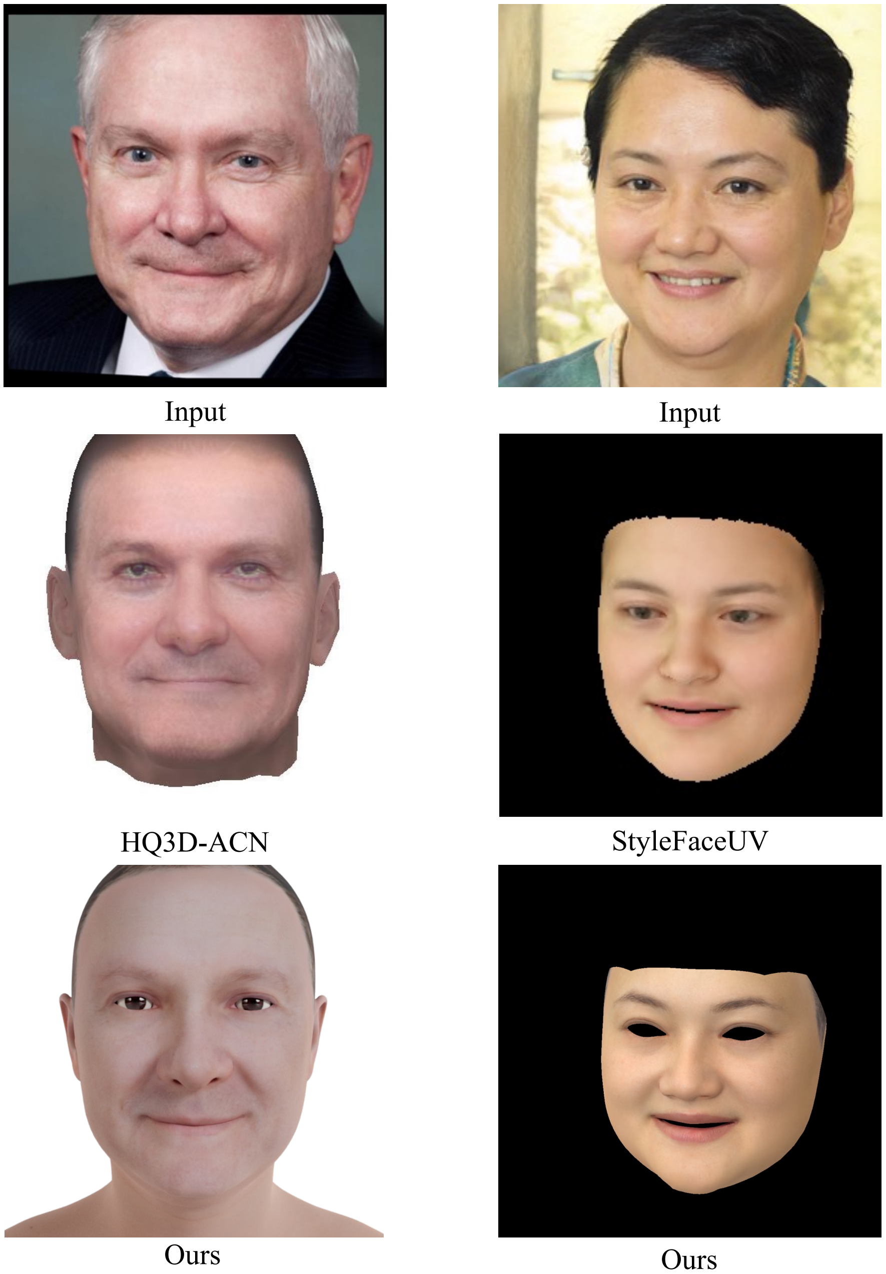

Figure 8. Visual comparisons of the reconstruction results to recent approaches HQ3D-ACN [7] and StyleFaceUV [4], where our results better resemble the input faces.

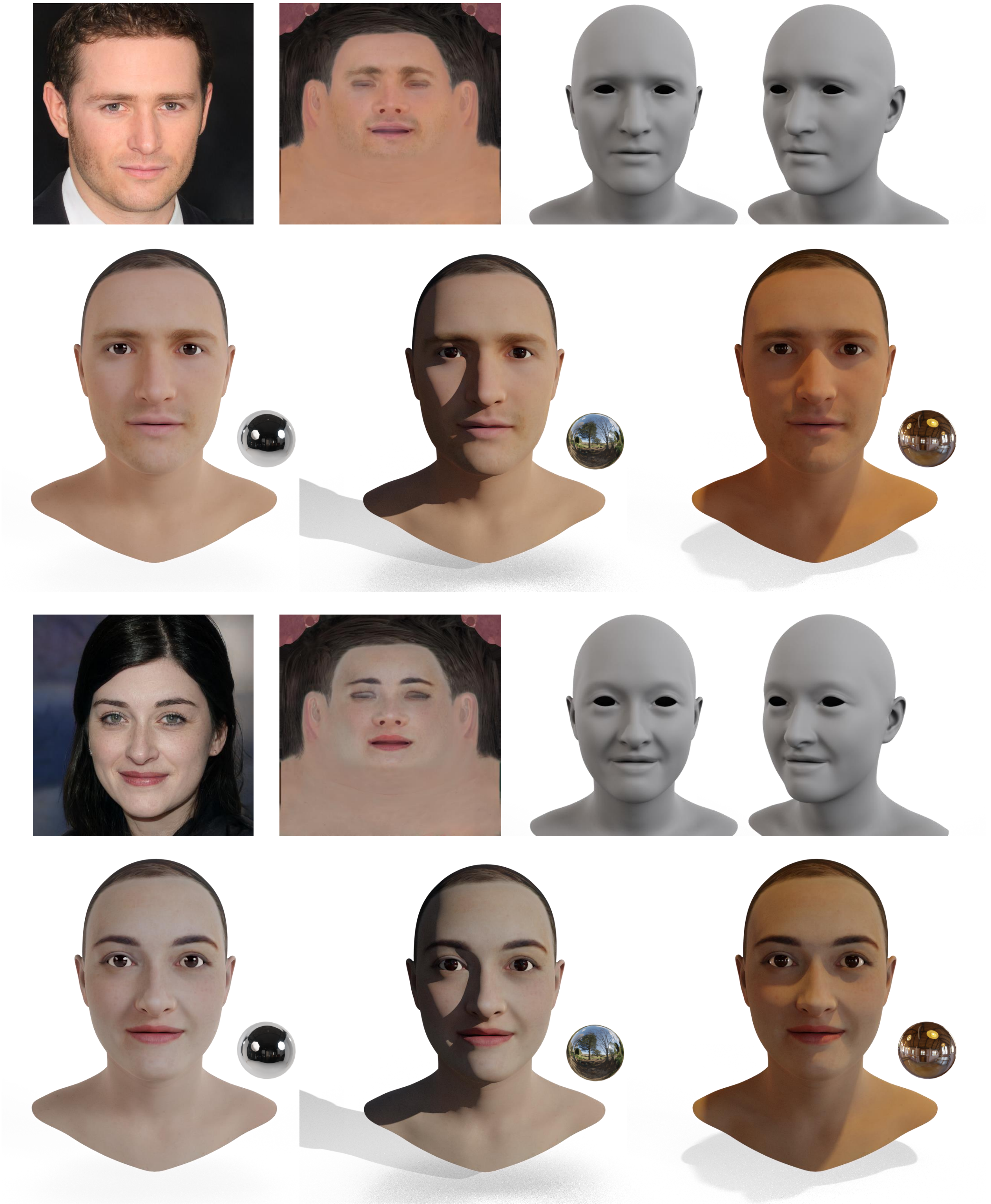

Figure 9. Examples of our reconstructed texture UV-maps, shapes, and renderings, where the produced textures are of high quality and without shadows which can be rendered with different lighting conditions.

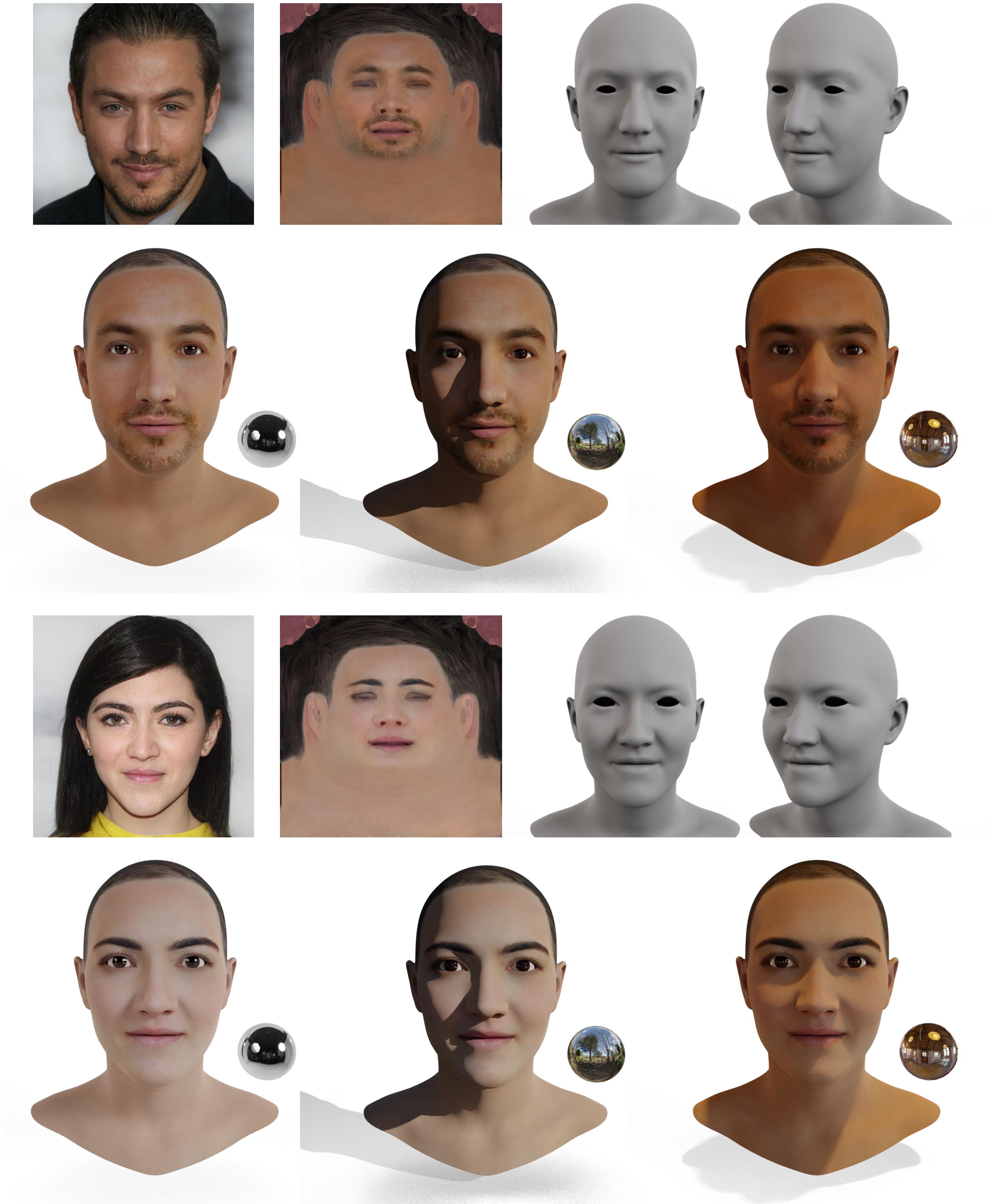

Figure 10. Examples of our reconstructed texture UV-maps, shapes, and renderings, where the produced textures are of high quality and without shadows which can be rendered with different lighting conditions.

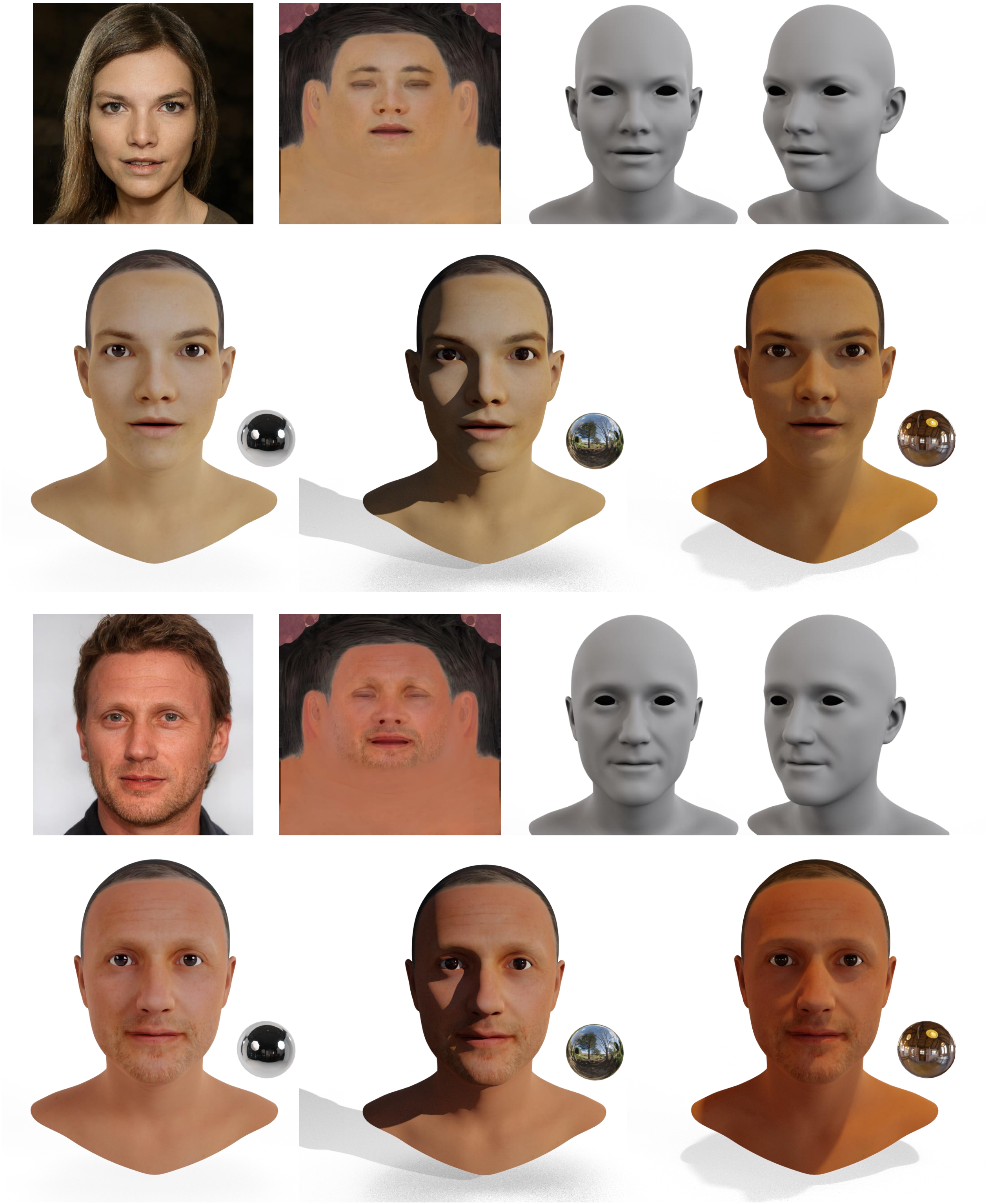

Figure 11. Examples of our reconstructed texture UV-maps, shapes, and renderings, where the produced textures are of high quality and without shadows which can be rendered with different lighting conditions.

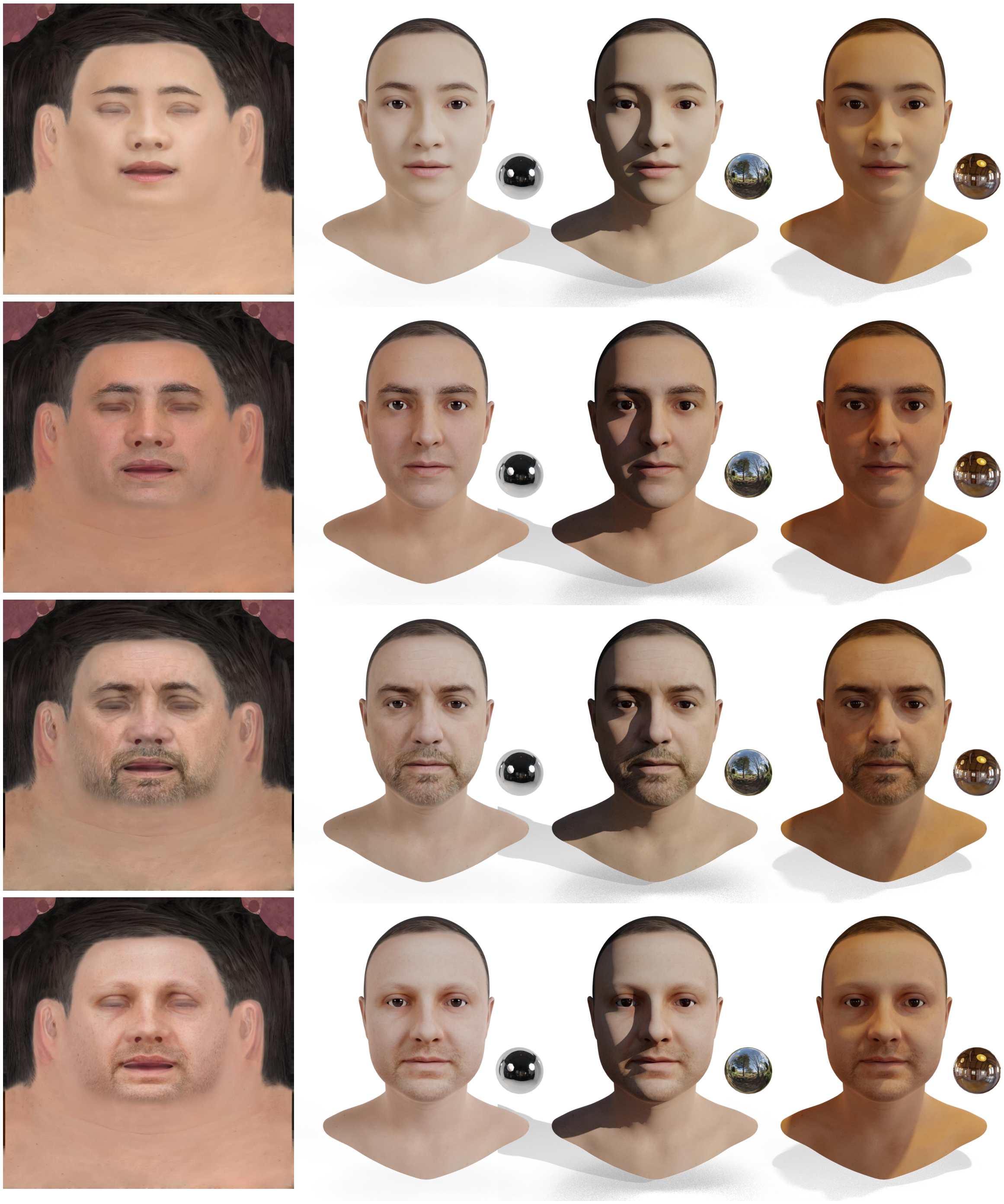

Figure 12. Examples of the proposed FFHQ-UV dataset, which are with even illuminations, neutral expressions, and cleaned facial regions (e.g., no eyeglasses and hair), and are ready for realistic renderings.

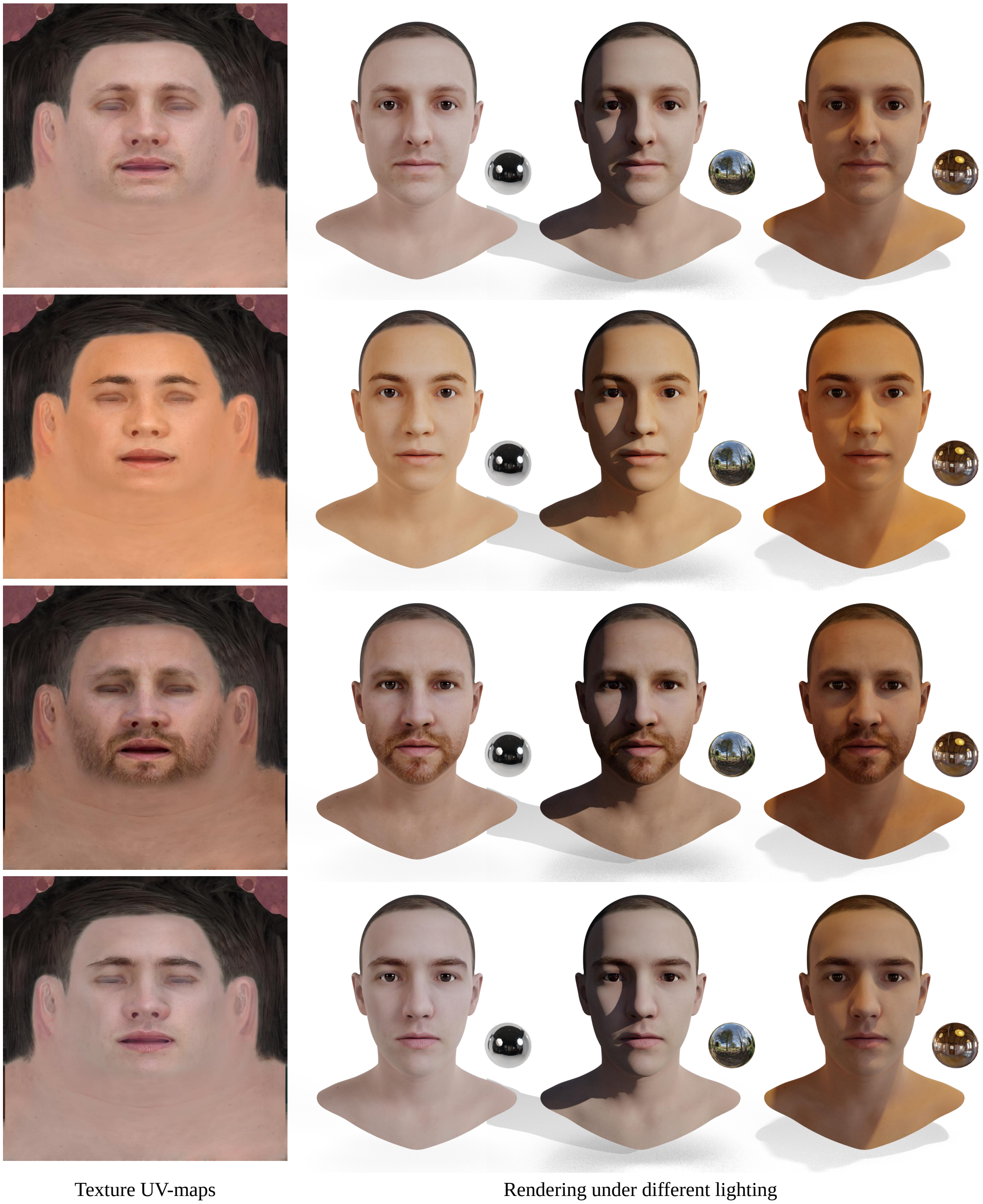

Figure 13. Examples of the proposed FFHQ-UV dataset, which are with even illuminations, neutral expressions, and cleaned facial regions (e.g., no eyeglasses and hair), and are ready for realistic renderings.