HUST: High-Fidelity Unbiased Skin Tone Estimation via Texture Quantization

Supplementary Material

In this supplementary material, we first provide a detailed description of the codebook training process (Sec. 1. Then we introduce the datasets used in the main paper (Sec. 2). Finally, we present additional qualitative results to demonstrate the robustness of our method (Sec. 3).

1. Codebook Learning Details

We further detailed the training loss functions in codebook learning. We use reconstruction loss, commitment loss, and adversarial loss. The reconstruction loss $\mathcal{L}_{rec}$ is defined as $\mathcal{L}_{rec} = \left\|\mathbf{I} - \hat{\mathbf{I}}\right\|^2$. Following the previous work [3, 7], we also adopt commitment loss to reduce the distance between the quantized feature $\mathbf{z}_q$ and the feature embedding $\mathbf{z}$:

$$\mathcal{L}_{com} = \|\mathrm{sg}\left[\mathcal{E}(\mathbf{I})\right] - \mathbf{z}_q\|_2^2 + \beta \left\|\mathrm{sg}\left[\mathbf{z}_q\right] - \mathcal{E}(\mathbf{I})\right\|_2^2, \quad (1)$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator and β is a weight factor. To further reduce the difference between the reconstructed textures and the input textures, we introduce an adversarial training procedure with a patch-based discriminator $\mathcal{D}$, and the adversarial loss $\mathcal{L}_{adv1}$ is defined as: $\mathcal{L}_{adv1} = \log(\mathcal{D}(\mathbf{I})) + \log(1 - \mathcal{D}(\hat{\mathbf{I}}))$. Finally, the overall optimization objective $\mathcal{L}_{pretrain}$ in the codebook pre-training process is:

$$\mathcal{L}_{pretrain} = \mathcal{L}_{rec} + \mathcal{L}_{com} + \lambda_0 \cdot \mathcal{L}_{adv1}, \quad (2)$$

where the loss weight λ_0 is set as 0.8.

2. Datasets

To create a high-quality facial texture codebook, we compile a large-scale high-quality face dataset of 3,392,957 images. Our dataset is derived from the large-scale open-source face dataset, WebFace260M [8]. All images have face detection bounding boxes larger than 512×512 pixels. Furthermore, we filtered out a high-resolution subset containing 501,605 images, each with a face detection bounding box exceeding 1024×1024 pixels. We provide example images in Fig. 1 and statistical results in Fig. 4. For the initial two training phases, we employed our high-resolution subset. In the final phase, we expanded to utilize the entire dataset. We will open-source these images to enable more researchers to advance related fields. Same as WebFace260M, our dataset and its subsets are for research purposes only.

To verify the impact of large-scale datasets, we conducted an ablation study on the dataset during the third training phase. The existing VGGFace2 [1] contains 9,131 IDs, but the image resolution is low and the quality is poor. Considering that our dataset far surpasses VGGFace2 in terms of identity diversity and resolution, we sampled 9,131 IDs from the high-resolution subset as a comparison subset. We conducted experiments on VGGFace2, the comparison subset, and the complete dataset for the third training phase, verifying the impact of increased training data resolution and number of identities on overall results. Since

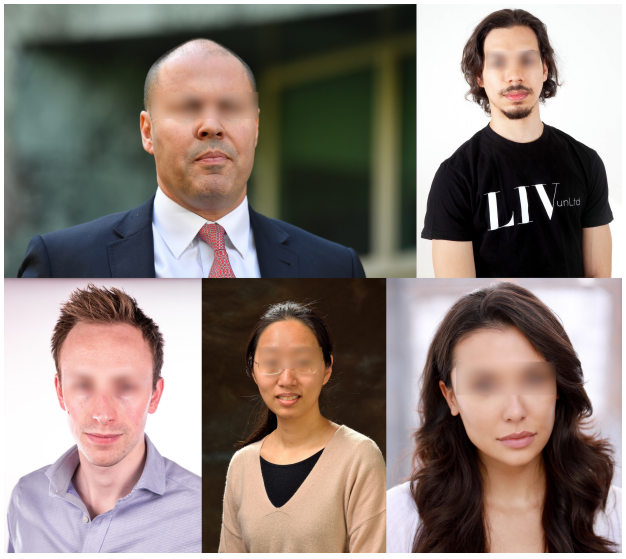


Figure 1. Examples of our large-scale face dataset, sourced from WebFace260M [8]. Our images contain a lot of skin detail on the face.

Method	IDs	PSNR↑	SSIM↑	LPIPS↓	ID↑
VGGFace2 [1]	9131	27.88	0.9060	0.1413	0.5870
Comparison Subset	9131	28.02	0.9096	0.1340	0.6118
Ours All	120,312	**29.14**	**0.9114**	**0.1109**	**0.7594**

Table 1. Ablation study of dataset. All metrics are computed on the CelebAMask-HQ dataset [5].

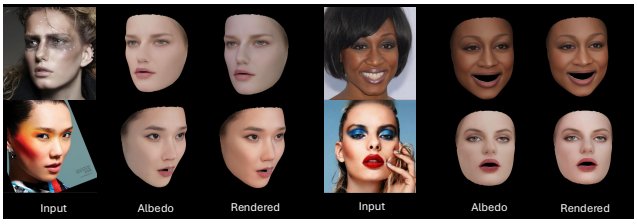


Figure 2. Examples of diffuse albedo reconstruction results for challenging inputs. We show the albedo results for the people wearing makeups.

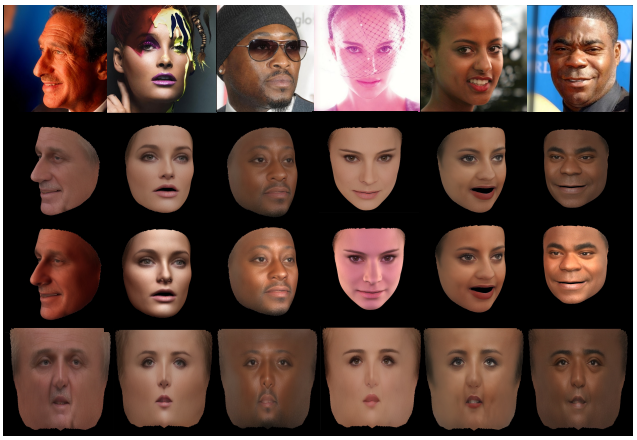
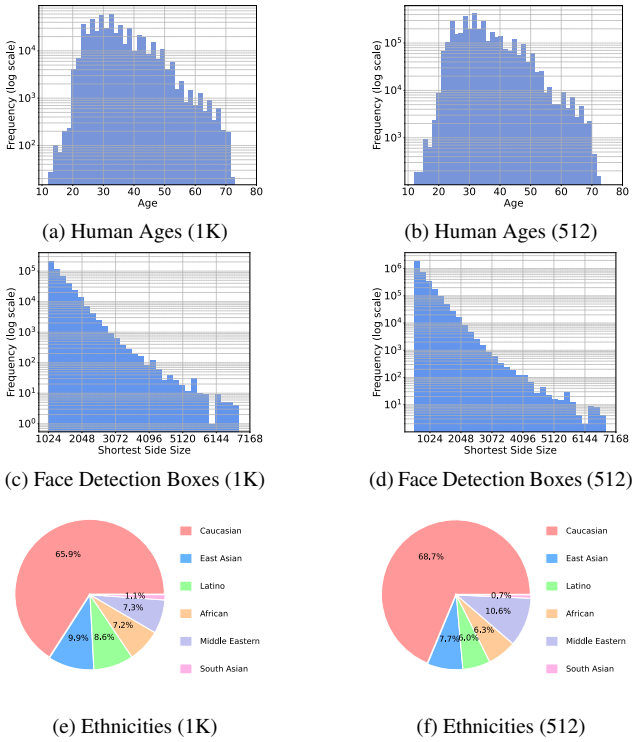


Figure 3. Examples of diffuse albedo reconstruction results for challenging inputs. The second, third, and fourth rows show the reconstructed albedo, reconstructed albedo with illumination, and UV albedo, respectively.



(a) Human Ages (1K)

(b) Human Ages (512)

(c) Face Detection Boxes (1K)

(d) Face Detection Boxes (512)

(e) Ethnicities (1K)

(f) Ethnicities (512)

Figure 4. Overview of our dataset. We show face size distribution age distribution and ethnicity distribution in the dataset, respectively.

CelebA-Mask-HQ [5] lacks ground truth (GT), we adopted the same approach as ID2Albedo [6]: overlaying the reconstructed albedo and illumination onto the original image to calculate metrics, with results shown in Tab. 1. The experiments demonstrate that improving resolution brings slight improvements to overall results, while further increasing the number of IDs significantly enhances model performance.

3. More Qualitative Results

We provide more albedo reconstruction results for challenging inputs (Fig. 2 and 3) and compare our methods with TRUST [4] and ID2Albedo [6]. The results in Fig. 5, 6 and 7 show that our diffuse albedo achieves the highest fidelity while maintaining comparable fairness. More details of the UV texture reconstruction are also shown in Fig. 8, 9, 10, 11 and 12.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 1

[2] 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 *CVPRW*, 2019. 5

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 1

[4] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *ECCV*, 2022. 2, 3, 4, 5

[5] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 1, 2

[6] Xingyu Ren, Jiankang Deng, Chao Ma, Yichao Yan, and Xiaokang Yang. Improving fairness in facial albedo estimation via visual-textual cues. In *CVPR*, 2023. 2, 3, 4, 5

[7] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 1

[8] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *CVPR*, 2021. 1, 3, 4

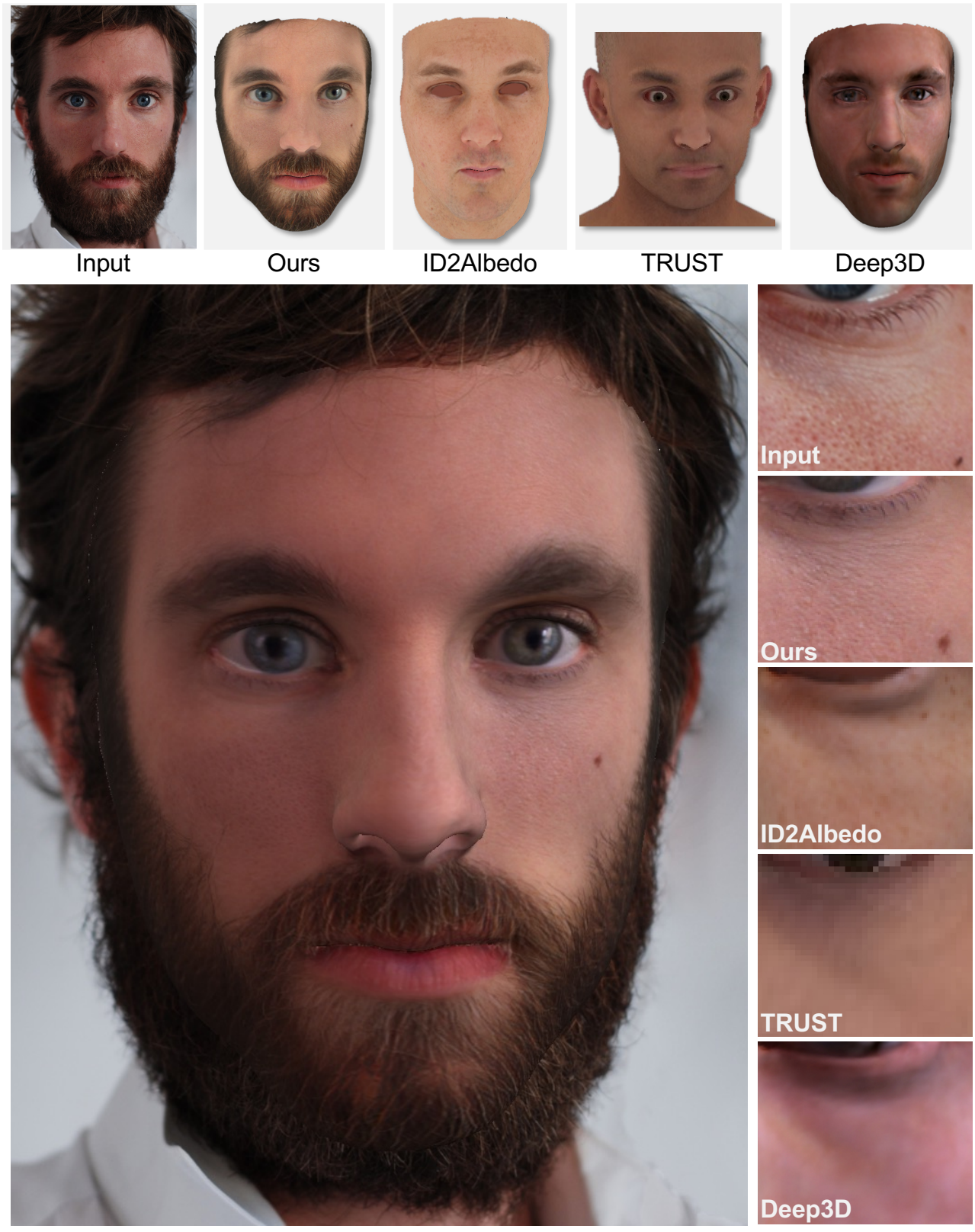


Figure 7. Comparisons on albedo details with ID2Albedo [6], TRUST [4] and Deep3D [2] diffuse albedo and rendered images. We achieve the most realistic rendered results and our generated high-fidelity albedo preserves the skin texture, pores, and moles on the face.

Figure 8. Visualization of texture reconstruction. Given a single image as input, our method reconstructs extremely high-quality texture details.

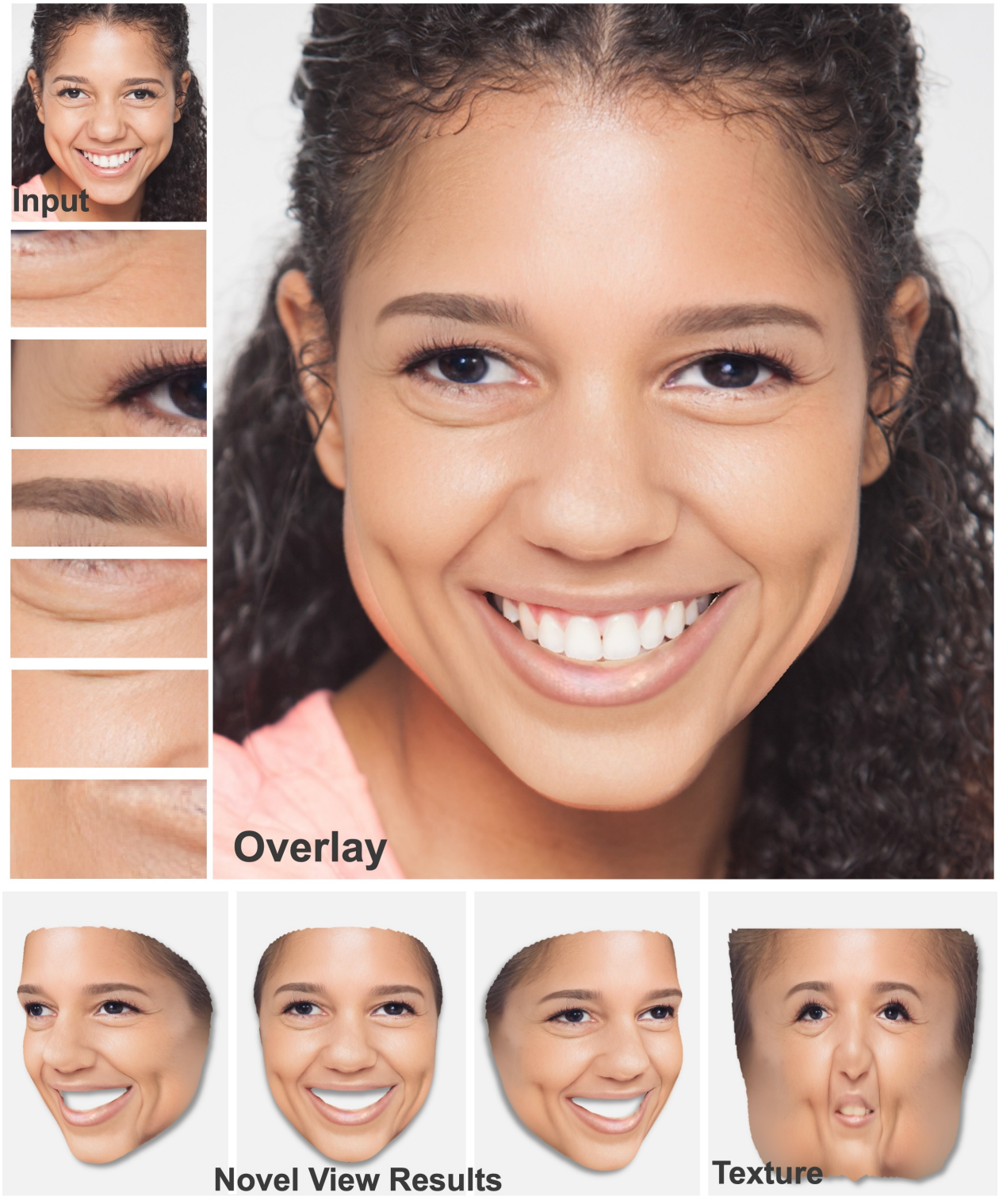


Figure 9. Visualization of texture reconstruction. Given a single image as input, our method reconstructs extremely high-quality texture details.

Figure 10. Visualization of texture reconstruction. Given a single image as input, our method reconstructs extremely high-quality texture details.

Figure 11. Visualization of texture reconstruction. Given a single image as input, our method reconstructs extremely high-quality texture details.

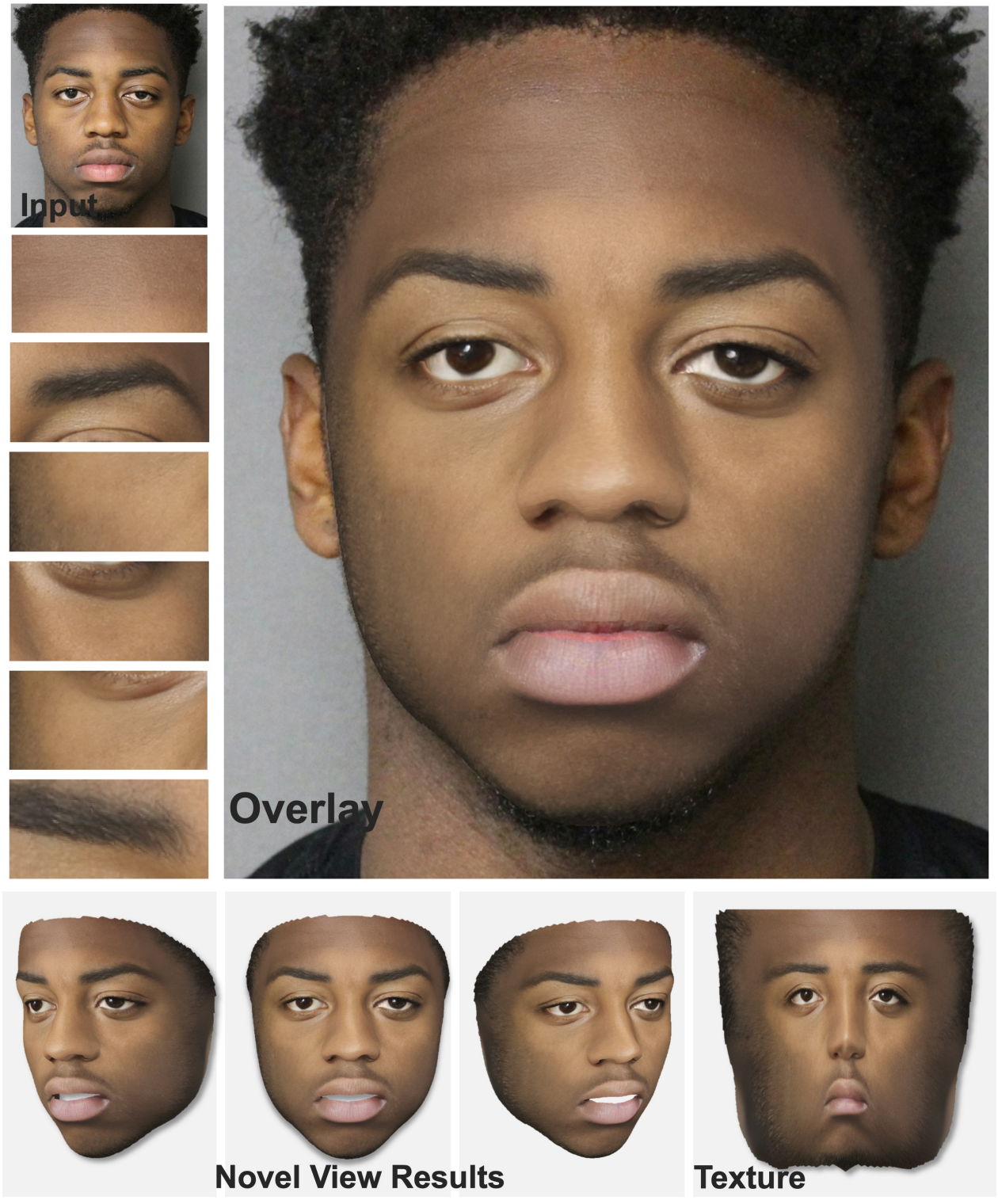


Figure 12. Visualization of texture reconstruction. Given a single image as input, our method reconstructs extremely high-quality texture details.