FreeUV: Ground-Truth-Free Realistic Facial UV Texture Recovery via Cross-Assembly Inference Strategy

Supplementary Material

In this supplementary document, we first describe the training data preparation process, including the generation of masked data and its use in both the in-the-wild and 3DMM domains (Appendix A). Then, we present ablation studies on network structures to analyze their impact on the 3DMM UV structure and realism (Appendix B). Finally, we provide additional qualitative comparisons with existing methods, showcasing the superior performance of our approach (Appendix D).

A. Training Data Preparation

Our training data preparation process is illustrated in Fig. S.1. Given a face image $\mathbf{I}$, we first derive the skin region mask $\mathbf{M}_I^w$ using a face segmentation algorithm [9]. Additionally, we extract another skin region mask $\mathbf{M}_I^m$ from the reconstructed 3DMM skin area. These two masks are combined via element-wise multiplication to produce the final mask $\mathbf{M}_I$, which represents the overlapping skin regions shared by both the in-the-wild image and the 3DMM-generated image. Applying this combined mask to the original image $\mathbf{I}$ yields the masked in-the-wild image $\mathbf{I}_w$. For both in-the-wild and 3DMM data, we utilize masked data to ensure consistency during the inference stage integration.

For 3D face reconstruction, we employ the FLAME 3DMM model [6]. We directly utilize the reconstruction method proposed in [7, 8] without additional training. The method in [7, 8] is a re-trained version of the Deep3Dface method [2], specifically adapted for the FLAME model. The reconstructed 3DMM result is denoted as $\hat{\mathbf{I}}_m$. Subsequently, we isolate the skin region from the UV texture by cropping and enlarging it, producing $\hat{\mathbf{T}}_m$. Using the reconstructed 3DMM shape, we sample pixels from the masked image $\mathbf{I}_w$ and unwrap them into UV space, resulting in the unwrapped UV texture $\mathbf{T}_w$. Similarly, the combined mask $\mathbf{M}_I$ is unwrapped to generate the UV mask $\mathbf{M}_T$. By applying the UV mask to the UV texture and UV position map from the 3DMM, we obtain the masked UV texture $\mathbf{T}_m$ and the UV position map $\mathbf{T}_{uv}$, respectively. These masked outputs are then reprojected into 2D space, resulting in the 2D images $\mathbf{I}_m$ and $\mathbf{I}_{uv}$.

This pipeline, starting from a single input image $\mathbf{I}$, produces a dataset suitable for training. However, it is notable that generating a complete and realistic UV texture is challenging. The accuracy of the unwrapped UV texture $\mathbf{T}_w$ is highly dependent on the performance of 3D face reconstruction. Flaws such as inaccuracies in 3DMM fitting and self-occlusion often result in distorted textures with missing regions.

Examples of the final in-the-wild images selected for training are shown in Fig. S.2. Images with failed face segmentation—such as those containing residual artifacts from hands or other objects—were manually excluded, as they do not accurately correspond to the skin region of the human face. During the inference stage, images with failed segmentation used as input can degrade the quality of the final output, as shown in Fig. S.3. To address this issue, we plan to leverage more advanced face segmentation techniques in the future to ensure robust performance in such cases. Nevertheless, our method is still capable of partially alleviating the impact of occlusions.

B. Additional Networks on Ablation Studies

Since both appearance network ϕ_a and structure network ϕ_s can independently generate UV texture, we evaluated their outputs to assess their ability to preserve the structure and realism of the 3DMM UV map. The results are shown in the 4th and 5th columns of Fig. S.4. Their inputs are the same as those of inference network ϕ_i, *i.e.*, the unwrapped UV texture $\mathbf{T}_w$ and the complete UV position map $\mathbf{\Upsilon}_w$. Neither of ϕ_a and ϕ_s is capable of preserving the structure of the 3DMM UV map. Structure network ϕ_s, trained on the 3DMM data domain, , fails to maintain consistency with the input UV map and introduces distortions.

We also observed that structural degradation still occurs when the structure network ϕ_s is trained using a UV-to-2D input-output configuration (see the 6th column of Fig. S.4), even though this setup aligns with the consistent pattern of the 3DMM data domain.

In contrast, FreeUV adopts a UV-to-UV Cross-Assembly Inference strategy, ensuring consistent alignment and structural integrity within the UV space. This strategy effectively allows the appearance network ϕ_a and the structure network ϕ_s to complement each other's strengths.

C. Comparison with DSD-GAN

To the best of our knowledge, DSD-GAN [3] is the first and only method to achieve complete UV texture completion without relying on ground truth UV data. It employs a dual-space discriminator GAN, applying two discriminators in both UV map space and image space to learn facial structures and texture details. Under the same UV ground-truth-free setting, our key contribution lies in improving

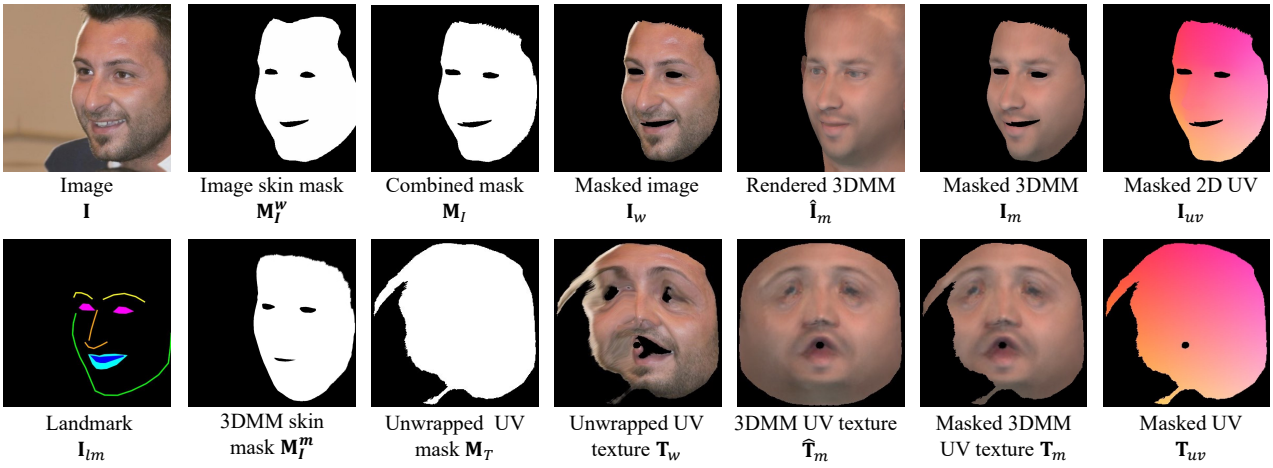

Figure S.1. **Training Data Preparation Process.** From a single image $\mathbf{I}$, this process produces a dataset suitable for training.

Figure S.2. **Examples of Final In-the-Wild Images Selected for Training.** Images with segmentation failures—such as those containing residual artifacts from hands or other objects—were excluded to ensure data quality.

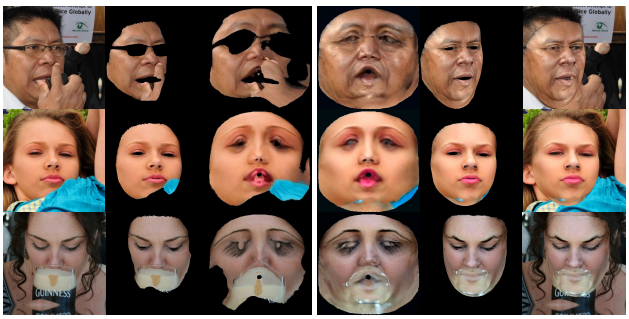

Figure S.3. **Impact of Segmentation Failures on Inference Results.** Failed segmentation in input images can degrade output quality. Advanced face segmentation techniques will be helpful in addressing this issue.

structural consistency and texture fidelity. Furthermore, our method leverages a pre-trained diffusion model, enhanc-

Figure S.4. **Ablation Studies on Network Structures.** Results using only the appearance network ϕ_a, only the shape network ϕ_s, and FreeUV with different configurations highlight the challenges in preserving the structure of the 3DMM UV map and maintaining realism.

Figure S.5. **Comparison with DSD-GAN.** Our method shows improved structural alignment and texture fidelity.

ing robustness to out-of-domain scenarios. As shown in Fig. S.5, DSD-GAN exhibits misalignment artifacts, particularly in the nasal and lip regions, as observed in their paper (since the official code is not publicly available).

D. Additional Results

We present additional results in Figs. S.6, S.7, and S.8. Compared to HRN [4], FFHQ-UV [1], and UV-IDM [5], our method excels in capturing fine details, achieving realistic outputs, and demonstrating enhanced robustness, particularly in preserving features such as beards, wrinkles, specular highlights, and makeup.

Figure S.6. **Comparison of Results.** Our method outperforms HRN [4], FFHQ-UV [1], and UV-IDM [5] in capturing fine details, achieving realism, and maintaining robustness.

Figure S.7. **Comparison of Results.** Our method outperforms HRN [4], FFHQ-UV [1], and UV-IDM [5] in capturing fine details, achieving realism, and maintaining robustness.

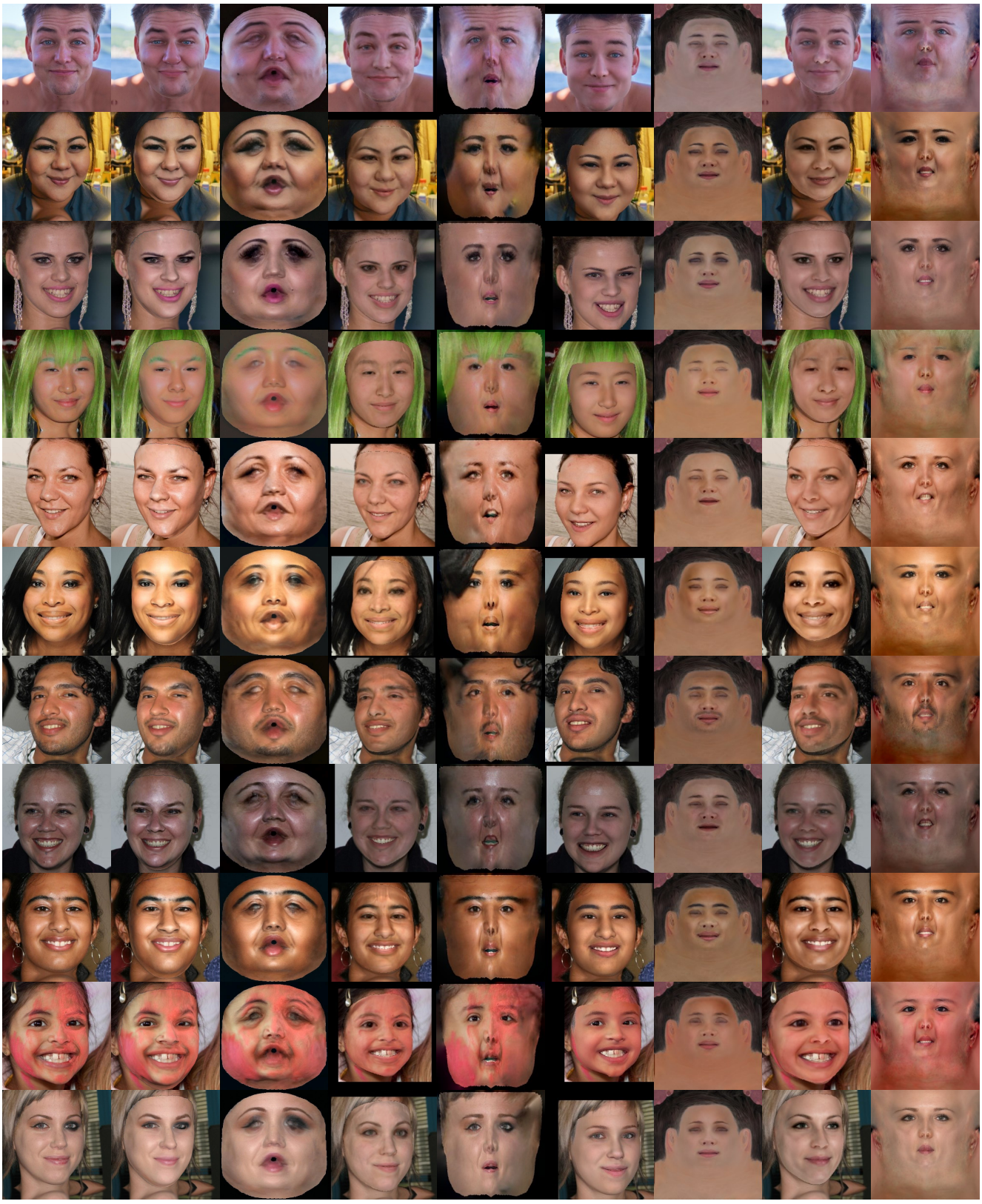

Figure S.8. **Comparison of Results.** Our method outperforms HRN [4], FFHQ-UV [1], and UV-IDM [5] in capturing fine details, achieving realism, and maintaining robustness.

References

[1] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. FFHQ-UV: Normalized facial uv-texture dataset for 3D face reconstruction. In *CVPR 2023*, pages 362–371, 2023. 2, 3, 4, 5

[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 *CVPR 2019 Workshops*, pages 285–295, 2019. 1

[3] Jongyoo Kim, Jiaolong Yang, and Xin Tong. Learning high-fidelity face texture completion without complete face texture. In *CVPR 2021*, pages 13970–13979, 2021. 1

[4] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *CVPR 2023*, pages 394–403, 2023. 2, 3, 4, 5

[5] Hong Li, Yutang Feng, Song Xue, Xuhui Liu, Bohan Zeng, Shanglin Li, Boyu Liu, Jianzhuang Liu, Shumin Han, and Baochang Zhang. UV-IDM: identity-Conditioned latent diffusion model for face UV-texture generation. In *CVPR 2024*, pages 10585–10595, 2024. 2, 3, 4, 5

[6] 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*, 36(6):194:1–194:17, 2017. 1

[7] Xingchao Yang, Takafumi Taketomi, and Yoshihiro Kanamori. Makeup extraction of 3D representation via illumination-aware image decomposition. *Computer Graphics Forum*, 42(2):293–307, 2023. 1

[8] Xingchao Yang, Takafumi Taketomi, Yuki Endo, and Yoshihiro Kanamori. Makeup prior models for 3D facial makeup estimation and applications. In *CVPR 2024*, pages 2165–2175, 2024. 1

[9] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV 2018*, pages 334–349, 2018. 1