UV-IDM: Identity-Conditioned Latent Diffusion Model for Face UV-Texture Generation

Supplementary Material

In this supplementary material, we first provide additional analysis of the proposed data creation pipeline (Sec. A). Afterward, we present more examples of the BFM-UV dataset (Sec. B). Then, we describe more implementation details of UV-IDM (Sec. C). Next, we show more visual results of extracting face UV textures from in-the-wild images using UV-IDM and rendering them, comparing with other state-of-the-art methods such as Deep3D [2], HRN [6], FFHQ-UV [1], OSTeC [4], NextFace [3] and AvatarMe [5]; and supplement with quantitative comparison results with HRM and Deep3D under less occluded frontal views (Sec. D). Finally, we discuss the social impact (Sec. E).

A. Additional Analysis of the Data Creation Pipeline

In Section 3.1 of the manuscript, we outline the creation process of the BFM-UV dataset (i.e., StyleGAN-Based Face Image Editing and Face UV-Texture Extraction). In this section, we further provide details on face UV-texture extraction for the dataset.

Deep3D, officially open-sourced by Microsoft, is built on BFM [7] and selects 35,709 vertices for optimization. However, these vertices cannot cover the entire UV texture map in the UV space. To ensure that the neural network can learn more easily, we use the original 53,215 vertices to guarantee the completeness of the UV texture map we create, which can fill the entire UV space. As shown in Fig. 2 of the manuscript, we use Deep3D to extract the predicted texture from the front-facing bald face and then feed it into the Linear Blending Module to obtain the final blended texture. Similar to the average texture used in FFHQ-UV for color adjustment and filling in the missing regions of the three-view texture, we present the visual comparison results of feeding the BFM-based average texture into the Linear Blending Module. As shown in Fig. A, using the textures predicted by Deep3D as a filling template, we can significantly alleviate the color discontinuity issue at the seam caused by different mask boundaries during the multi-view incomplete UV fusion process. Compared to using average texture filling, the predicted UV textures with our method demonstrate more realistic effects in hair and beard regions.

B. More Examples of BFM-UV Dataset

In Section 3.1.2 of the manuscript, we show some examples of the created BFM-UV dataset. In this supplementary material, we further provide more visual results in Fig. B.

C. More Implementation Details

In this section, we further describe the training and inference details of UV-IDM. During training, inspired by [8], UV-IDM's training consists of two phases: training a VAE to compress the UV pixel space into a semantically richer, low-dimensional latent space, then freezing the VAE's encoder and jointly training ICM and LDM in this space. Within ICM, only the embedding network τ_θ is trainable, encoding incomplete UV textures into conditional embeddings to guide LDM's denoising. Incomplete textures are extracted using pre-trained Deep3D and BiSeNet, along with predefined UV mapping, see Section 3.1.2 of the manuscript. Additionally, τ_θ is custom-designed, lightweight, and initialized with random weights. The inference process requires ICM, LDM, and VAE's decoder. ICM extracts and encodes an incomplete texture from a raw image I into conditional embedding. Together with the denoising time condition t, they guide LDM from noise to generate latent variables consistent with I, which VAE's decoder uses to generate the full texture.

D. More Results of UV Textures Generation

In this section, we first present the visual results of our method in terms of multi-view consistency. Specifically, we select multiple images of the same face from different viewpoints, use one viewpoint for texture prediction, and then render the predicted results to compare with the original images from other viewpoints, as shown in Fig. C. Subsequently, we present more visual comparisons of face UV textures generated by different methods, including our UV-IDM combined with Deep3D and HRN (resulting in D-UV-IDM and H-UV-IDM), as well as Deep3D, HRN, FFHQ-UV, OSTeC, NextFace, and AvatarMe. Among them, AvatarMe provides visual effects in both texture space and 3D space.

As shown in Fig. D, we display the rendering results and UV texture maps for various methods. It can be observed that compared to Deep3D and FFHQ-UV, D-UV-IDM is able to accentuate more realistic textures, such as the man's beard in the third group and the woman's cheeks in the fourth group. Notably, FFHQ-UV tends to yield smoother outcomes, aligning with the findings depicted in the manuscript. In contrast to HRN (the number of optimization iterations is set to 0), H-UV-IDM demonstrates strong robustness to occlusions while preserving authentic textures, exemplified by the women's hair in the first and

fourth groups, the man's glasses in the third group, and the woman's head adornments in the fifth group. Moreover, compared to AvatarMe, we can achieve noticeably more realistic rendering results that are closer to the actual image, as shown in Fig. E.

We present additional texture generation results of UV-IDM, HRN (the number of optimization iterations is set to 50), OSTeC, and NextFace under more challenging circumstances in Fig. F. HRN and NextFace produce textures that suffer from substantial artifacts, which can be attributed to the interplay between shape and texture optimization during the iterative fitting process, as indicated by the microphone in the third row, the backgrounds in both the fourth and seventh rows and the sunglasses in the fifth row. UV-IDM, on the other hand, adapts well to these scenarios. Compared to OSTeC, the textures generated by UV-IDM are replete with finer details, even from occluded angles. Take, for instance, the woman in the eighth row: OSTeC renders disparate details on the left and right sides of her face, whereas UV-IDM ensures that the generated textures under large pose variations remain reasonable. These indicate that our method can generate more symmetrical, detailed, and robust textures against occlusions.

We test 650 and 750 less occluded frontal images from FFHQ and CelebAMask-HQ to further demonstrate our ability to generate high-fidelity textures. The results are shown in Table A. Compared to HRN and Deep3D, UV-IDM achieves advantages when provided with less occluded frontal images.

E. Social Impact

This work primarily focuses on enhancing the texture fidelity of 3D face reconstruction, which can contribute positively to various applications such as virtual reality, gaming, and digital content creation. By improving the quality of 3D assets, it can lead to more immersive and realistic experiences in virtual environments. However, it is essential to consider the potential negative social impacts that may arise from the misuse of such generative techniques. For instance, the improved 3D face reconstruction methods could be exploited to create fake videos or images, which can be used for spreading misinformation, identity theft, or other malicious purposes. Additionally, concerns about privacy and consent may arise when using individuals' facial data without their permission. Therefore, while advancing technological developments, it is essential to carefully consider how to regulate and manage the use of these techniques to prevent potential misuse and their adverse impacts.

References

[1] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. *arXiv:2211.13874*, 2022. 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 *CVPR Workshops*, 2020. 1

[3] A. Dib, G. Bharaj, J. Ahn, C. Thébault, P. Gosselin, M. Romeo, and L. Chevallier. Practical face reconstruction via differentiable ray tracing. *Comput. Graph. Forum*, 2021. 1

[4] Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Ostec: One-shot texture completion. In *CVPR*, 2021. 1

[5] 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. 1

[6] 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. *arXiv:2302.14434*, 2023. 1

[7] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, 2009. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

Table A. Quantitative comparison of rendering quality of less occluded frontal facial views.

Method	FFHQ			CelebAMask-HQ		
	LPIPS↓	FID↓	CSIM↑	LPIPS↓	FID↓	CSIM↑
HRN	0.1360	70.88	0.9689	0.1312	62.02	**0.9713**
H-UV-IDM	**0.1131**	**54.00**	**0.9717**	**0.1159**	**53.89**	0.9677
Deep3D	0.1306	58.24	0.9578	0.1338	59.89	0.9552
D-UV-IDM	**0.1085**	**52.47**	**0.9710**	**0.1138**	**52.61**	**0.9674**

Figure A. Visualization results of blended textures generated by feeding the BFM average texture and the Deep3D predicted texture into the Linear Blending Module, respectively.

Figure B. Triplets in the BFM-UV dataset. The first three columns show in-the-wild images from three different viewpoints. The middle three columns show the corresponding hair-removed bald portraits from the same three viewpoints, and the last column displays the UV texture maps that we extract.

Figure C. The textures (Left-UV, Front-UV, and Right-UV) are generated using the multi-view images (Left, Front, Right) separately. Then, each texture is rendered back to the three views. "Left Render Front" refers to the result of rendering the texture generated from the left view image to the front view image. Other similar terms are explained similarly. Our method can generate a unified texture map for multiple views.

Figure D. Additional visualization results. Every two rows form a group, with the first row showing the rendering results of different methods onto the original image and the second row displaying the corresponding UV texture maps for each method.

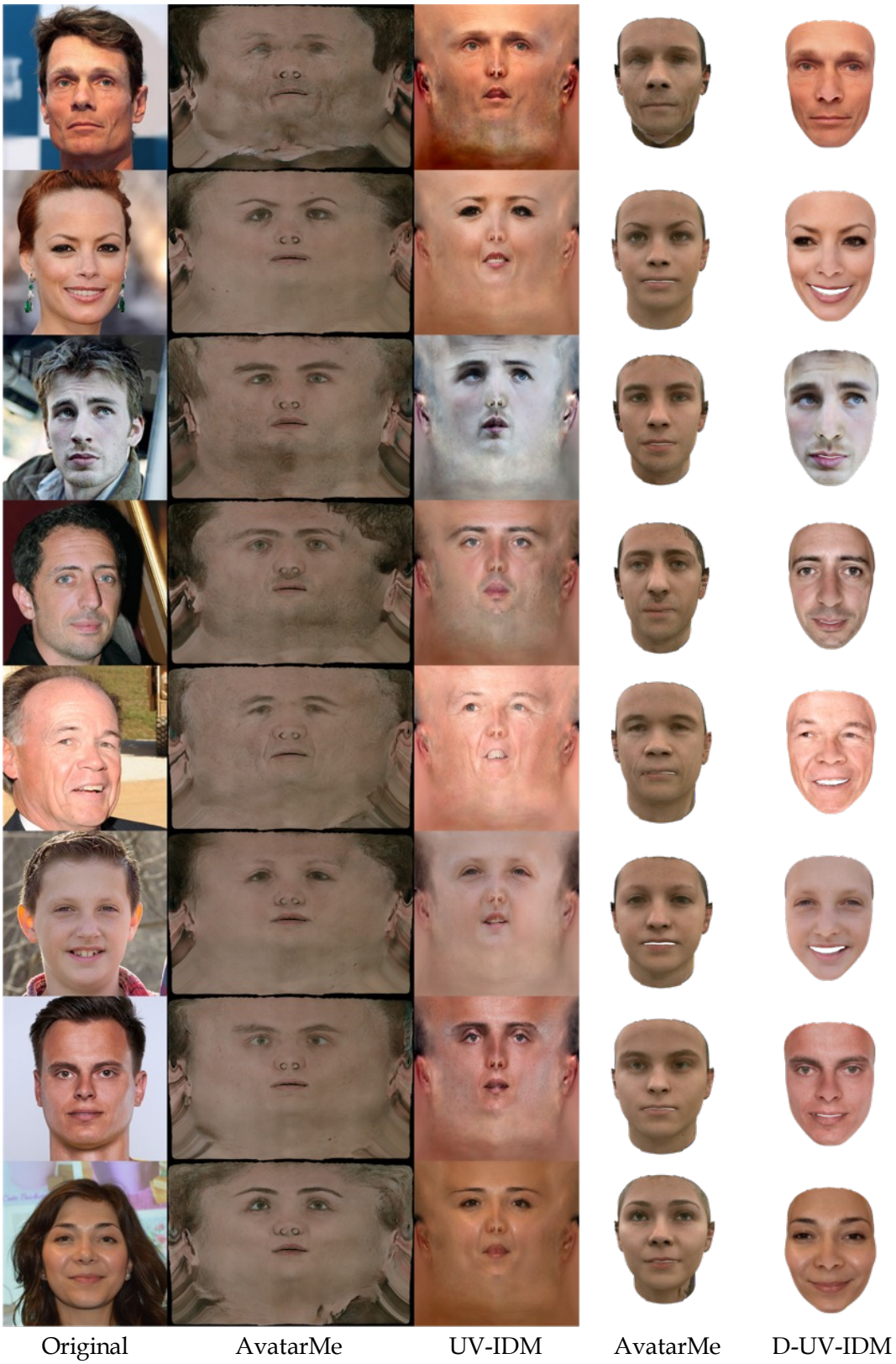

Figure E. Visual comparison of texture generation and 3D reconstruction between UV-IDM/D-UV-IDM and AvatarMe.

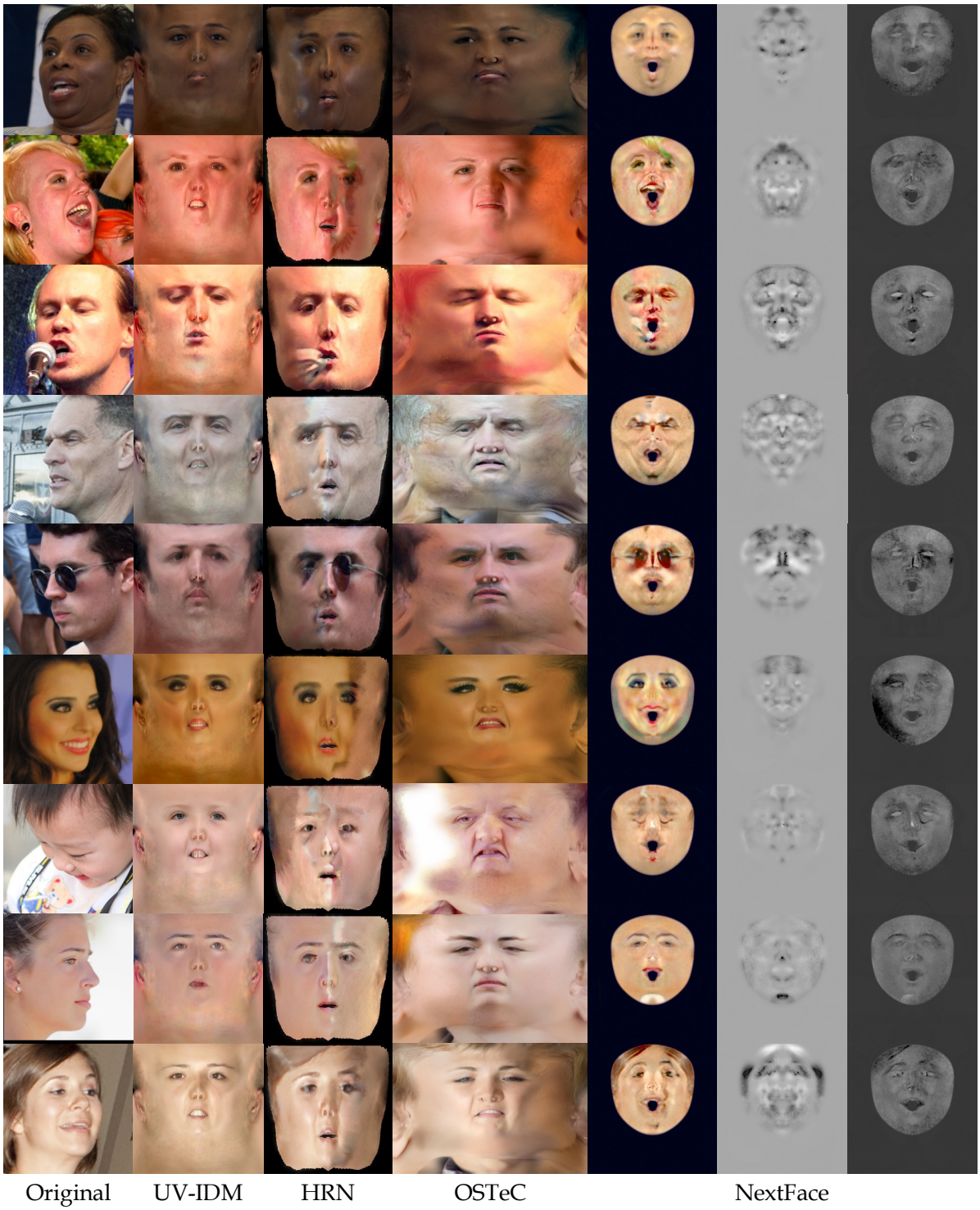

Figure F. Visual comparison of textures generated by UV-IDM, HRN, OSTeC, and NextFace.