High-Fidelity 3D-aware GAN Inversion by Pseudo-multi-view Optimization
Supplementary materials

Jiaxin Xie*[1] Hao Ouyang*[1] Jingtan Piao †[2] Chenyang Lei[3] Qifeng Chen†[1]
[1]HKUST [2]MMLab, CUHK [3] CAIR, HKISI-CAS

A. Implementation details

A.1. Optimization settings

We utilize a pretrained EG3D model trained on the FFHQ [5] dataset for optimization. The model utilizes the triplane with 256 resolution and generates 3D-aware photorealistic images at 512 resolution.

Regarding the camera pose, we utilize a readily available face reconstruction network [3] to obtain p_0, as per the method described in [1]. Prior to utilizing [3], we need to align the input image with the pose distribution of the pretrained EG3D, as illustrated in Figure 1. The code for pose estimation can be found in our Github repository.

For the initial visibility estimation stage, we train our model with a learning rate of 5e-3 and 1000 iterations. The optimization process included 500 iterations for the latent code optimization in the $\mathcal{W}+$ space and 500 iterations for generator fine-tuning, We update all learnable parameters in the EG3D generator to obtain the tuned model. However, the mesh obtained directly from EG3D has misalignment with the input image. Therefore, we reconstruct the mesh from the depth rendered from the same view as the input image. After obtaining the mesh, we calculate the visibility for each vertex by utilizing the z-buffer of the rasterization algorithm to determine whether other mesh faces occluded this vertex. Then, we warp the texture of the input image to visible vertices and rasterize the mesh to obtain novel view masks $\mathcal{M}_v$ and visible textures $\mathcal{V}_i$, thereby completing the visible part reconstruction. For the occluded part, we use the generator to generate it. To blend the visible and occluded parts, we experimented with two methods to blur the boundary: directly applying a Gaussian kernel with a radius of 10 or using Poisson blending. We generated pseudo-multi-views by blending visible and occluded parts with the smoothed boundary. Finally, for the optimization stage, we set the learning rate at 3e-4 and the training iteration at 3000 using the synthesized pseudo-multi-views and input image.

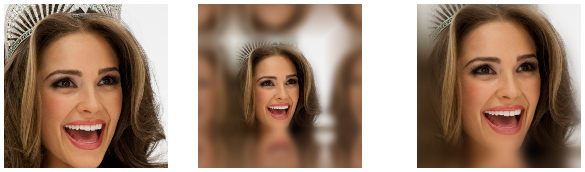

Original Aligned with padding Aligned cropped input

Figure 1. Visualization of input image alignment.

A.2. Image attribute editing

For image attribute editing, we follow the pipeline proposed in [7] to calculate the latent direction. We first generate 500,000 images with the canonical view. We then predict the attributes of the synthesized images and rank them by scores. We use the 10,000 samples with the highest and 10,000 with the lowest scores. We then randomly use 70% of them for training a linear SVM and 30% for testing the accuracy of the trained classifier. We calculate the attribute direction with the trained SVM.

B. Analysis

B.1. More quantitative evaluation metrics

B.1.1 IBRNet

We utilized the 3D consistency evaluation setting introduced in [4], employing the pretrained IBRNet [9] model to predict the input view from five synthesized novel views. The five views were carefully selected by choosing the canonical view and the four edge views on the sphere camera trajectory within a yaw range of [-0.35, 0.35] rad and a pitch range of [-0.25, 0.25] rad. Our main paper presents the PSNR metric that evaluates the difference between the ground truth images and the reconstructed images generated by IBRNet. Additionally, we include the SSIM and lpips metrics in Table 1, demonstrating that our proposed method surpasses other baseline models.

B.1.2 Pose accuracy

We adopt the pose accuracy metric from [1]. With the face reconstruction model [3], we randomly select one novel

*Joint first authors
†Joint corresponding authors

view for every CelebA-HQ test image and estimate the pose for the novel view outputs of all the methods. We compute the L2 loss between the estimated pose and the GT pose that is used to render the novel view. From table 1, we find that the rendered novel view images of all three methods keep good pose consistency with the GT pose, the error is minor, and the margin between different methods is extremely small.

B.1.3 Identity

An intuitive criterion for judging the 3D-aware GAN inversion is whether it keeps the identity of the input person in novel views. We have evaluated the criterion with a user study in the main paper, asking people which video keeps the best identity of the input image. We also provide quantitative metrics. We randomly select one novel view for every CelebA-HQ test image and compute the mean Arcface [2] cosine similarity between rendered novel view images and input images. The comparison is shown in table 1. Our ID loss exceeds other baselines by a large margin.

B.2. Texture-geometry trade-off

As analyzed in Figure 2 of the main paper, the quality of the synthesized view severely decreases as the optimization iteration increases. We also provide quantitative metrics as shown in Figure 2. We calculate the 3D consistency metric using the IBRNet [9] for every 500 optimization steps. As the PSNR of the reconstruction increases, the 3D consistency metrics, on the contrary, decreases. The quantitative metrics also indicate a degradation in novel view quality as the optimization process continues, which matches the findings in our qualitative analysis.

C. Additional visual results

We provide more visual results, including the failure cases and the qualitative comparison with baselines.

C.1. Failure cases

In Figure 3, we demonstrate two failure cases for our approach. The first source image contains hands, and the estimated geometry is incorrect. We can see obvious blurry regions in the synthesized novel views. The second case contains the out-of-distribution pose with trinkets. The generated face suffers from slight distortion, and the trinkets' shape is incorrect.

C.2. Qualitative comparison

More qualitative comparison are shown in Figure. 5, Figure. 6 and Figure. 7.

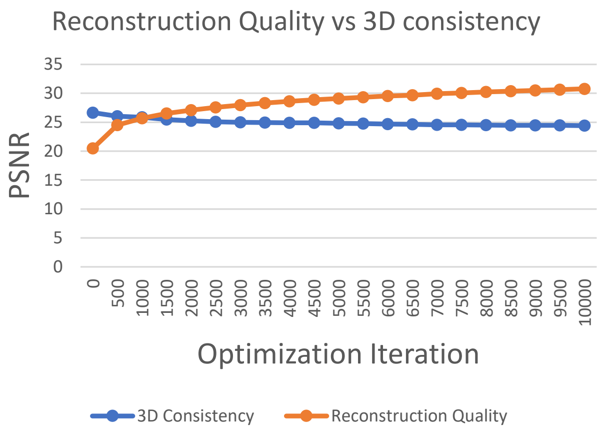
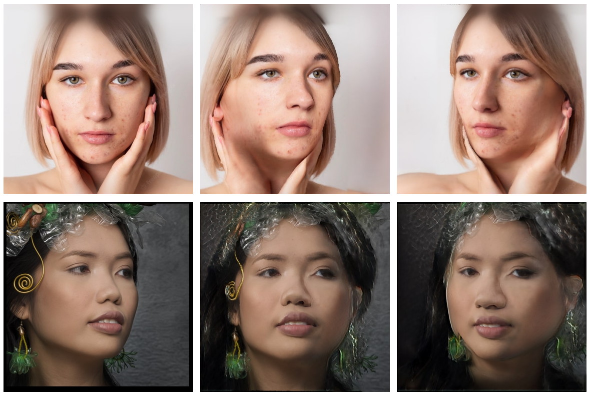


Figure 2. As the reconstruction quality increases with the optimization iteration, the 3D consistency on the contrary decreases.



Input image Novel view 1 Novel view 2

Figure 3. Failure cases.

Method	PSNR↑	SSIM↑	Lpips↓	Pose ↓	ID ↑
PTI [6]	21.20	0.697	0.457	0.04178	0.657
IDE-3D [8]	20.69	0.676	0.462	**0.04152**	0.671
Ours	**21.69**	**0.734**	**0.429**	0.04179	**0.744**

Table 1. More quantitative evaluation metric on 3D consistency.

D. Alternative choices

An alternative regularization strategy to improve the geometry is to add regularization on density while the reconstruction loss is still calculated on the single input. With an initially estimated geometry, we can regularize the density during the training. We add an additional loss which requires the density of the current output is similar to the correct geometry. However, although it helps to keep the geometry, the synthesized novel view contains blurry details. As in Figure 4, compared to density regularization, our pseudo-multi-view generates clearer details and keeps higher fidelity. The possible reason is that the single in-

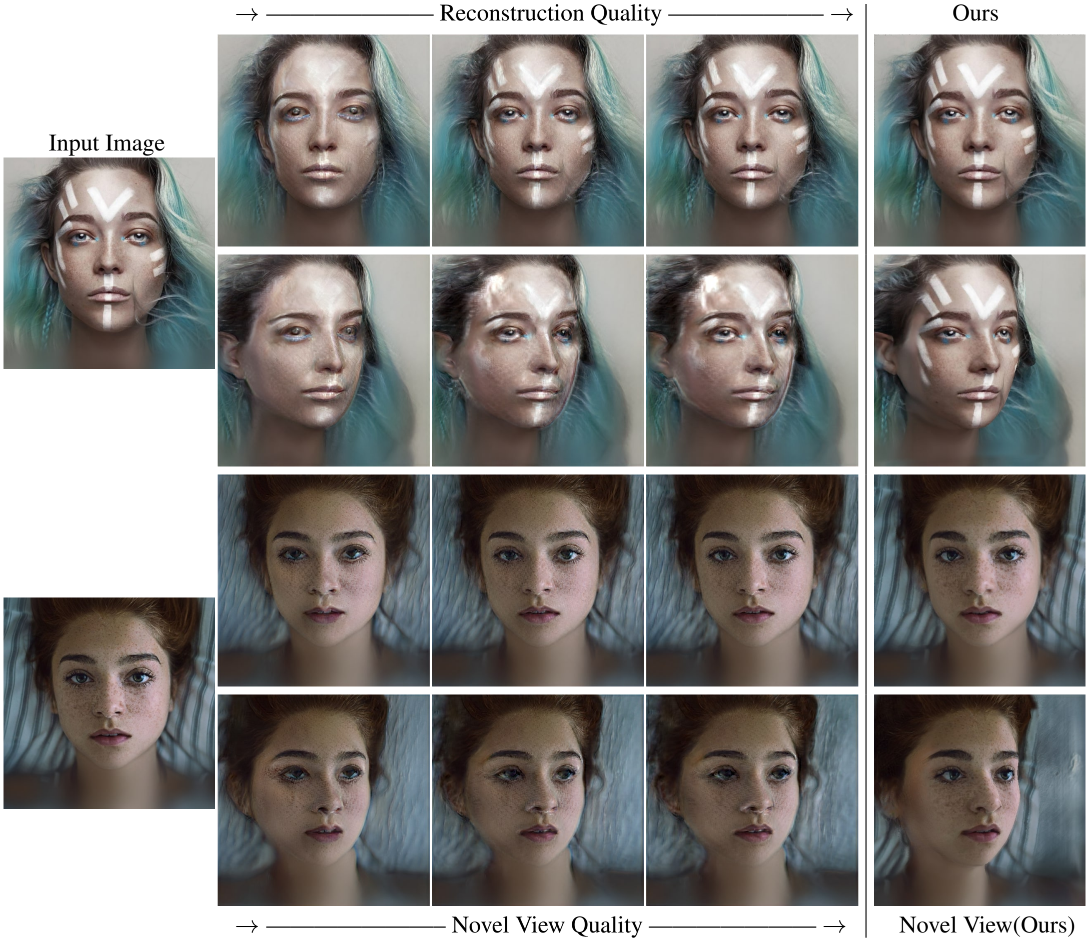


Figure 4. Reconstruction quality vs. novel view quality during the optimization process using density regularization. Although the geometry of the synthesized novel view will not distort like in Figure 2 in the main paper, the generated textures contain visually-unpleasant noisy details compared to our results. Zoom for details

put as supervision contains not enough information for constructing photo-realistic details. The pseudo-multi-view can better solve the ambiguity.

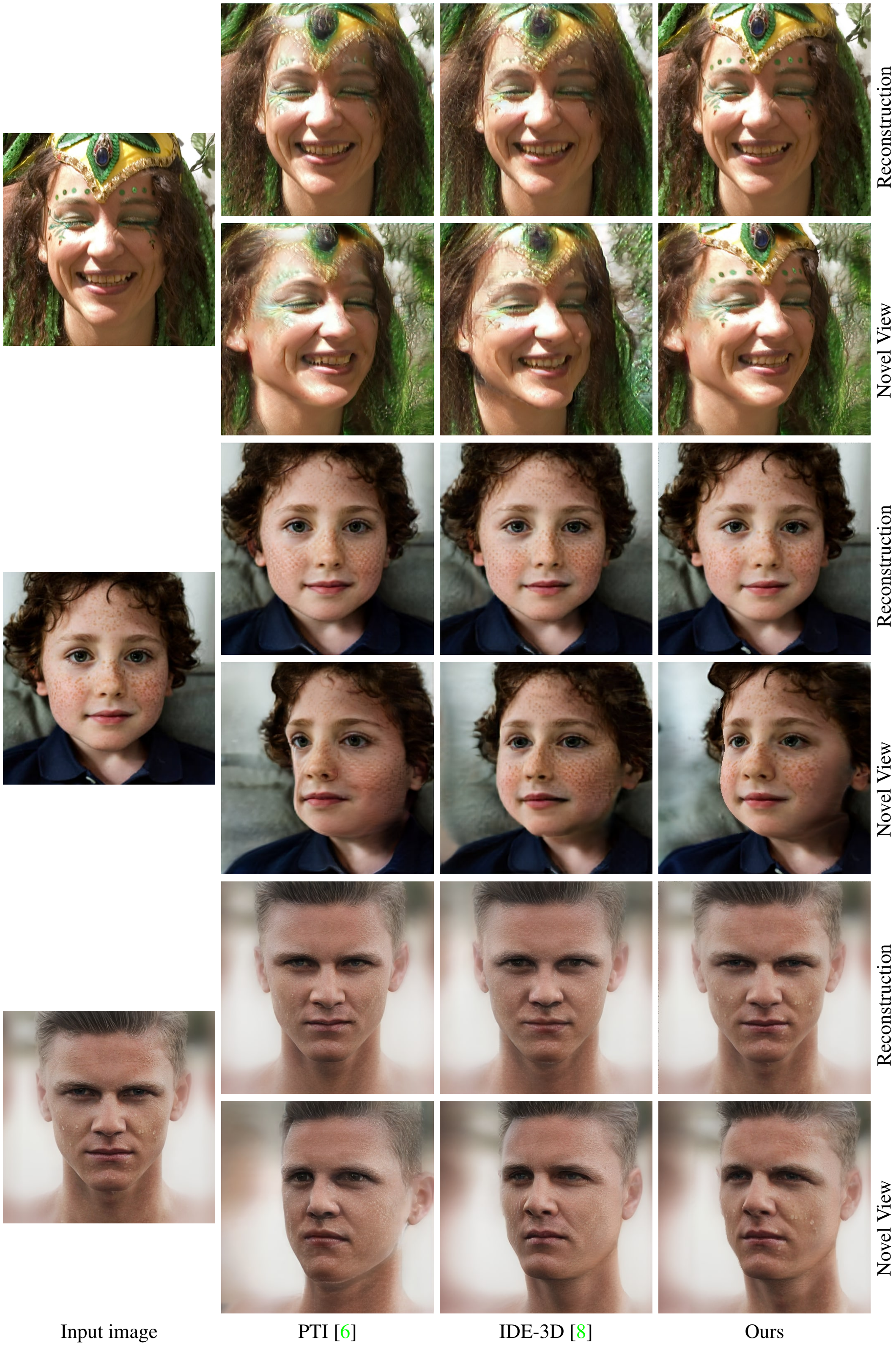


Figure 5. **More qualitative comparison with baselines.**

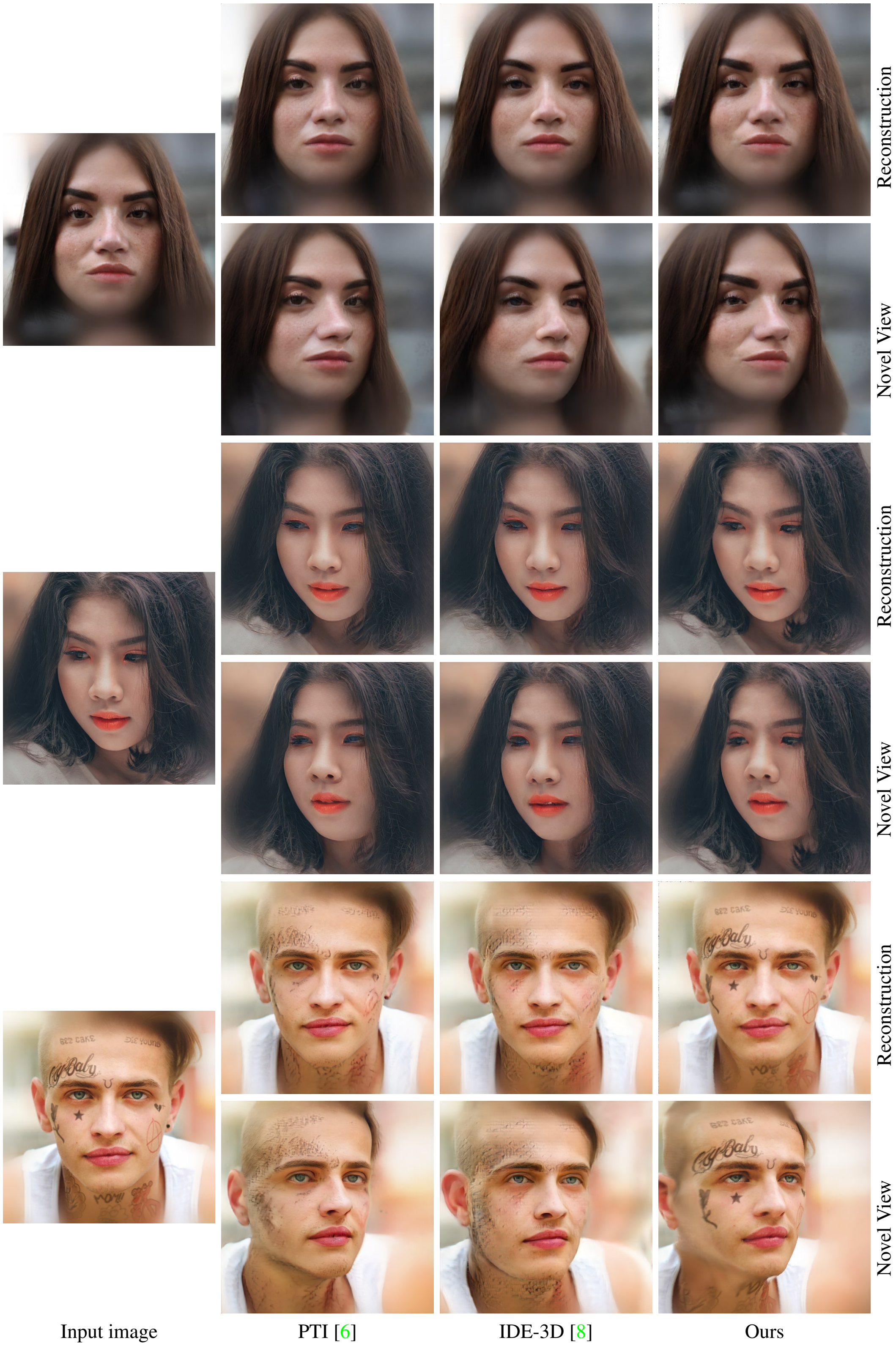


Figure 6. **More qualitative comparison with baselines.**

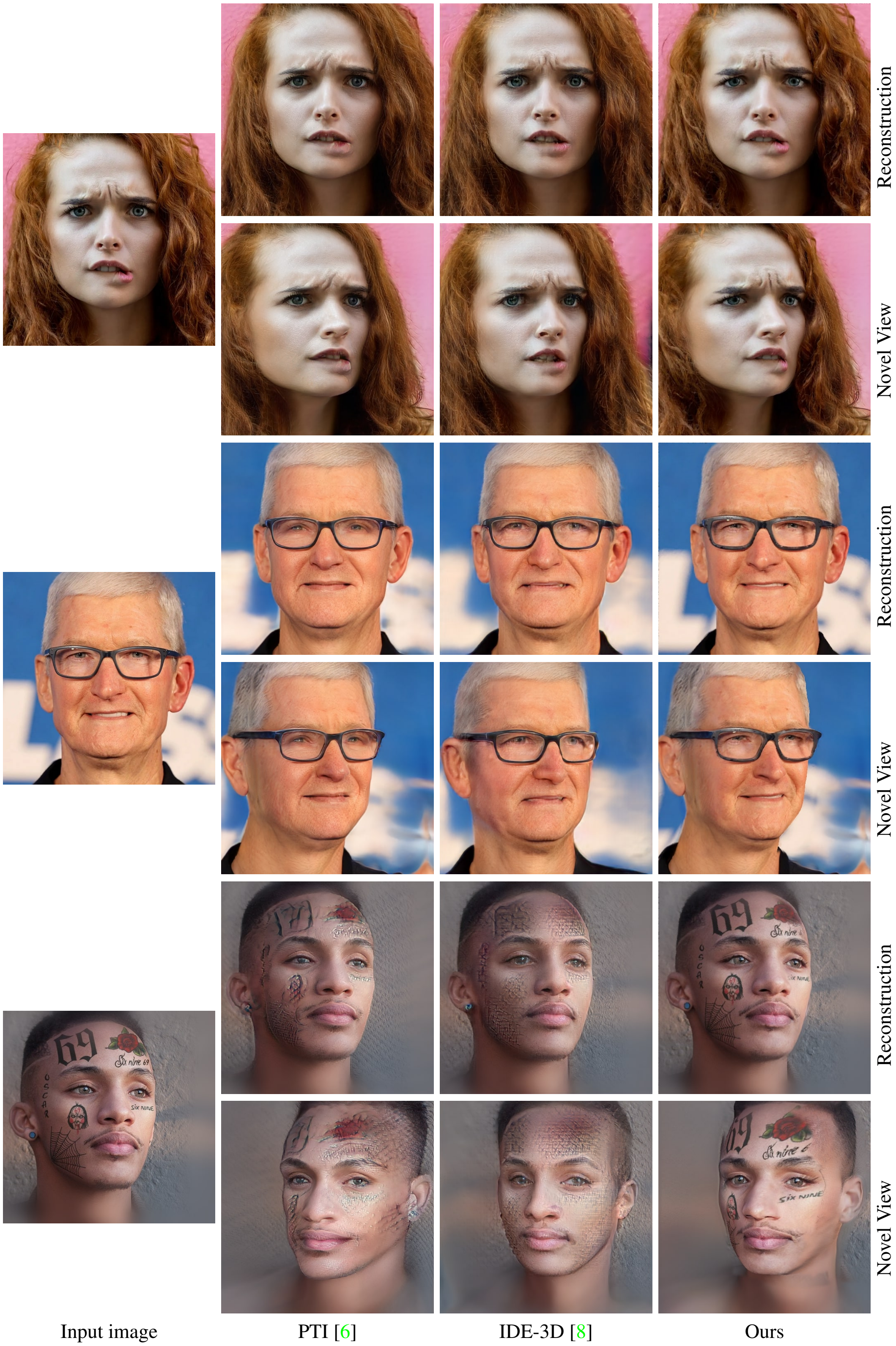


Figure 7. **More qualitative comparison with baselines.**

References

[1] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 1

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[4] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 1

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, 2019. 1

[6] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 2, 4, 5, 6

[7] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. 1

[8] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *arXiv preprint arXiv:2205.15517*, 2022. 2, 4, 5, 6

[9] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 1, 2