Supplemental Material
PREIM3D: 3D Consistent Precise Image Attribute Editing from a Single Image in Real Time

In the supplement, we first provide implementation details, including encoder training process and edit directions seeking. We follow with additional experiments and visual results. We highly recommend watching the supplemental video, which contains a live demonstration of the real-time inversion and attribute editing and a demonstration of sequential editing synthesis.

A. Implementation Details

A.1. Encoder Training

We implemented our encoder training on top of the official pSp [4] encoder training framework implementation. We set the $\lambda_{l2} = 1.0$, $\lambda_{lpips} = 0.8$, and $\lambda_{ori} = 0.4$ in the first $20,000$ training steps. After the $20,000$ steps, we gradually add a delta for the $\lambda_w = 1e^{-4}$ every $5,000$ steps. After the $100,000$ steps, we gradually add a delta for the $\lambda_{sur} = 1e^{-4}$ every $5,000$ steps. The in-domain images are sampled from yaw angles between $[-30°, 30°]$ and pitch angles between $[-20°, 20°]$. The surrounding images are sampled from yaw angles between $[-20°, 20°]$ and pitch angles between $[-5°, 5°]$

A.2. Edit Directions Seeking

We use InterfaceGAN [5] to train a SVM to find out the attribute editing directions. For the editing directions in the original space, the generator is applied to produce $140,000$ images. For the editing directions in the inversion manifold, we perform inversion with our encoder on FFHQ [3] dataset. Here, we have obtained the latent code w and image pairs. An off-the-shelf multi-label classifier based on ResNet50 [2] is applied to predict the images. We train the SVM (https://github.com/clementapa/CelebFaces_Attributes_Classification/) to find the hyperplane that distinguishes binary attributes using the latent code w and the corresponding classification result as input. The normal vector of the hyperplane is the attribute editing direction.

B. Comparison on Face Inversion at More Camera Poses

We uniformly sample 20 inverted images for each image of the first 300 images from CelebA-HQ in different yaws ranges using IDE-3D, 3D-Inv, Pixel2NeRF, and PREIM3D. As with the main text, IDE-3D and 3D-Inv perform image inversion with 500 w optimization steps and 100 generator fine-tuning steps. We show the identity consistency (ID) in Figure 1.

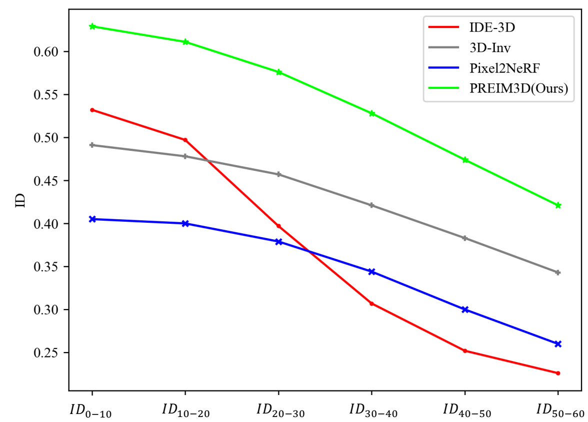

Figure 1. ID_{a-b} denotes the mean ArcFace similarity score between the input image and the 20 inverted images uniformly sampled from yaw angles between $[-b°, a°] \cup [a°, b°]$ and pitch angles between $[-20°, 20°]$. Our method has a higher ID score than other methods in different yaw ranges.

C. Additional Precise Editing

AA & AD. Following [7], we use attribute altering (AA) to evaluate the change of the desired attribute and attribute dependency (AD) to measure the degree of change on other attributes when modifying one attribute. AA is the change on the logit Δl_t of the off-the-shelf multi-label classifier detecting attribute t and is normalized by $\sigma(l_t)$, which is the standard deviation calculated from the logits of CelebA-HQ dataset. AD measures the change of logit Δl_i for other attributes $\forall i \in \mathcal{A} \setminus t$, where $\mathcal{A}$ is the set of all attributes. Here, we use the mean-AD, defined as $\mathbb{E}(\frac{1}{k}\sum_{i\in\mathcal{A}\setminus t}(\frac{\Delta l_t}{\sigma(l_i)}))$.

To further validate the precision of the editing in the inversion manifold, we perform more attribute editing. We make different degrees of editing by adjusting α, and then observe the changes on the other attributes. Figure 2, 3 shows the difference between editing in the original space and editing in the inversion manifold, involving goatee, lipstick gray hair, wavy hair, and gender attributes. Both 2D-space and 3D-space attribute editing show more precise editing in the inversion manifold than in the original space.

D. Naive Optimization-based Inversion

Different from the PTI technique, the naive optimization-based inversion method only optimizes the latent code w, while fixing the generator. Figure 4 shows the inversion results of the naive optimization-based inversion method.

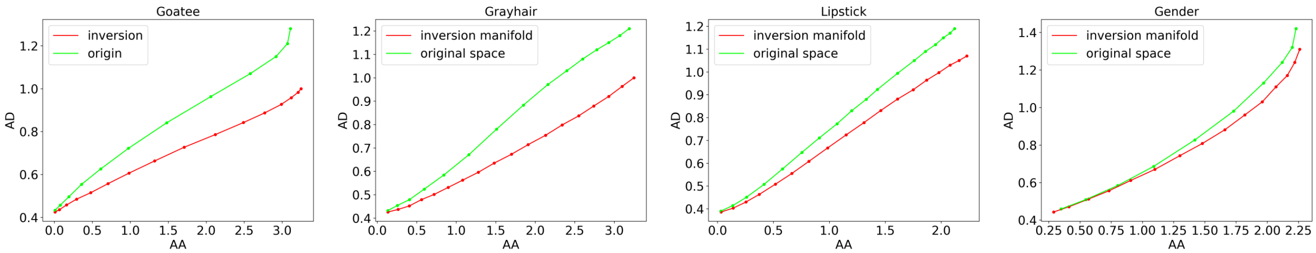

Figure 2. PREIMD(Ours). As the degree of editing α changes, both Attribute Altering (AA) and Attribute Dependency (AD) change. Lower AD indicates more precise.

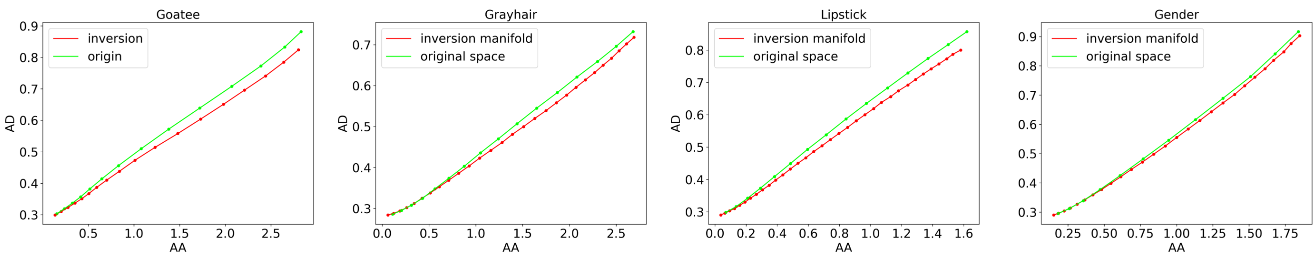

Figure 3. e4e(2D) [6]. As the degree of editing α changes, both Attribute Altering (AA) and Attribute Dependency (AD) change. Lower AD indicates more precise.

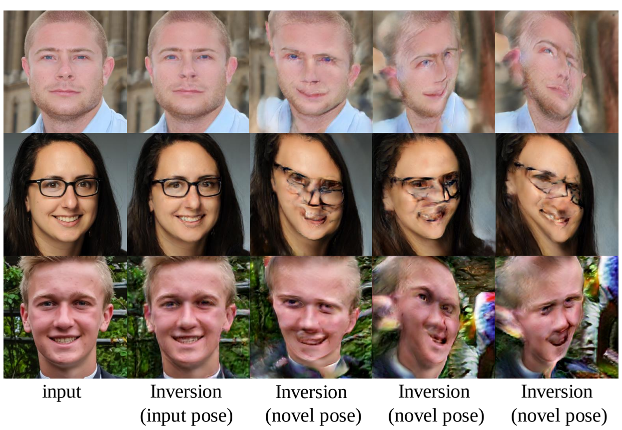

Figure 4. The inversion result of 1, 0000 iterations of steps. The naive optimization-based inversion method reconstructs the view of the input camera pose but produces significant artifacts in the views of other camera poses.

Figure 5. It is complex to train the encoder and fine-tune the generator at the same time. We found some tough ripple-like artifacts in the hair.

Figure 6. Inversion and (hair color) editing results on cat faces.

E. Fine-tuning the Generator

Inspired by Pixel2NeRF [1], we attempted to fine-tune the generator when training the inversion encoder. Unfortunately, there are always some ripple-like artifacts in the hair, which was also observed for Pixel2NeRF, as shown in the figure 5.

F. Beyond Human Face

We conducted some experiments with the AFHQ Cat. We invert the dataset to obtain inversion latent samples. Following GANSpace, We adopt principal component analysis (PCA) to find the semantic directions. The results in the cat domain are shown in Fig 6.

G. FID and KID

Method	FID_{ori}	FID_{sm}	FID_{mid}	FID_{la}	KID_{la}
IDE-3D	**22.7**	**36.8**	45.2	75.7	0.065
3D-Inv	28.1	40.6	**44.9**	65.4	0.046
Pixel2NeRF	83.3	85.4	86.2	93.2	0.086
PREIM3D (Ours)	43.6	48.3	50.7	**63.3**	**0.042**

Table 1. FID & KID comparisons on 1,000 faces from CelebA-HQ. FID_{ori} is measured between the inverted images at the original angle and the input images. We use sm, mid, la for uniform samples from yaw $[15°, 20°]$ and pitch $[10°, 15°]$, yaw $[25°, 30°]$ and pitch $[15°, 20°]$, yaw $[35°, 40°]$ and pitch $[20°, 25°]$.

We evaluated inversion FID in Table 1. The inception features used in FID focus on the whole image, while our method introduces regularization of the face regions, which makes our FID scores not as good as IDE-3D and 3D-Inv at small angles. However, our model outperforms previous works at large angles. KID shows similar results.

H. Additional Visual Results

We provide a large number of inversion and editing results produced by PREIM3D in Figure 7 to 13

References

[1] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3981–3990, 2022. 2

[2] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1510–1519. IEEE Computer Society, 2017. 1

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[4] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 1

[5] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[6] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2

[7] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 1

Figure 7. The inversion results obtained by PREIM3D. The first column is the input image.

Figure 8. The age editing results obtained by PREIM3D. The first column is the input image.

Figure 9. The eyeglasses editing results obtained by PREIM3D. The first column is the input image.

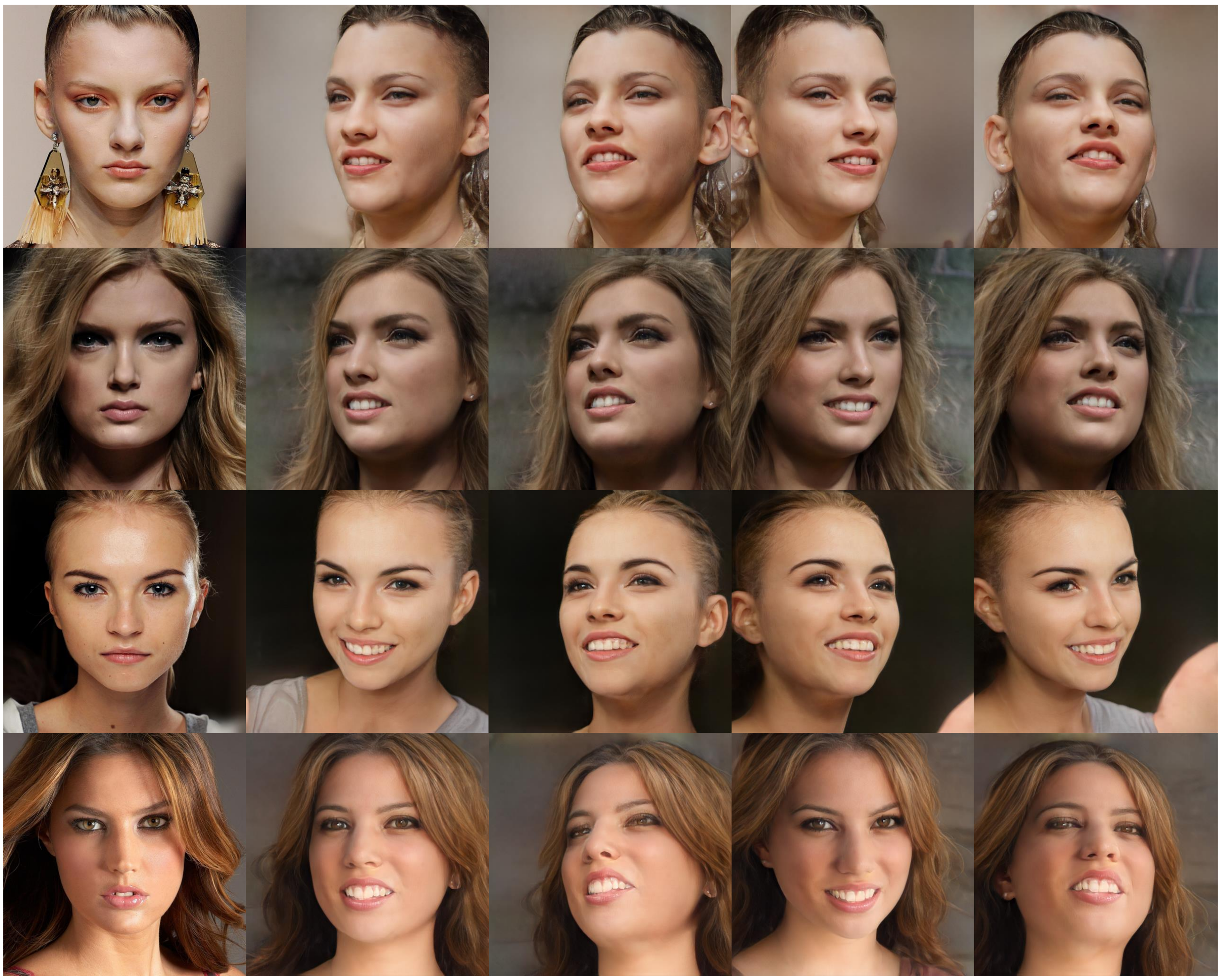

Figure 10. The smile editing results obtained by PREIM3D. The first column is the input image.

Figure 11. The goatee editing results obtained by PREIM3D. The first column is the input image.

Figure 12. The lipstick editing results obtained by PREIM3D. The first column is the input image.

Figure 13. The wavy hair editing results obtained by PREIM3D. The first column is the input image.