Appendix for High-Fidelity GAN Inversion for Image Attribute Editing

Tengfei Wang[1] Yong Zhang[2] Yanbo Fan[2] Jue Wang[2] Qifeng Chen[1]
[1]The Hong Kong University of Science and Technology [2]Tencent AI Lab

In Section A, we provide more details of the architecture design and training settings. In Section B, we give more details of the Shannon lower-bound and the information bottleneck hypothesis. In Section C, we demonstrate additional visual results of diverse attribute editing on images and videos.

A. Implementation Details

A.1. Training Details

We use the pretrained StyleGAN2 [5] as the generator, and e4e [12] as the basic encoder E_0 in all our experiments. Both generator and the basic encoder are held fixed during the training. We adopt Adam optimizer [6] with LookAhead technique [13]. The learning rate is set to $1e-4$. The iteration number is set to 90,000 with a batch size of 8. λ_{align} is set to 0.1, λ_{per} is set to 0.8, and λ_{adv} is set to 0.01. λ_{id} is set to 0.1 for face domain and 0.5 for car domain. The self-supervised training scheme for ADA is illustrated in Fig. 1.

A.2. Networks Details

C(out_c**,** s**)** denotes a Conv2d layer with output channel number out_c and stride s.

D(out_c**,** s**)** is composed of a bilinear upsampling layer followed by the ConvBlock.

ResBlock(out_c**)** is composed of two conv layers with PReLU activation.

Adaptive Distortion Alignment Module. It can be represented as: C(16,1)-C(32,2)-ResBlock(32)-C(48,2)-ResBlock(48)-C(64,2)-ResBlock(64)-D(64,2)-ResBlock(64)-D(48,2)-ResBlock(48)-D(32,2)-ResBlock(32)-C(3,1). Skip-connection is adopted for encoder layers (ie, C) and decoder layers (ie, D) with the same value of out_c.

Consultation Encoder. It can be represented as: C(32,1)-C(48,2)-ResBlock(48)-ResBlock(48)-C(64,2)-ResBlock(64)-ResBlock(64).

Fusion Layer. Fig. 2 compares Eq. (1-3) and StyleGAN2. Specifically, the original layers from StyleGAN are held fixed, and the latent map C_i from the consultation encoder is used to modulate the feature maps.

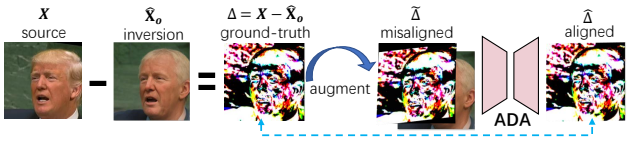

Figure 1. Self-supervised training of ADA.

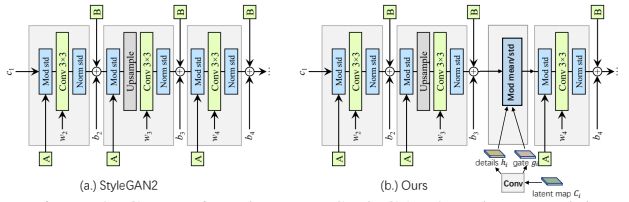

Figure 2. Comparison between StyleGAN2 and our model.

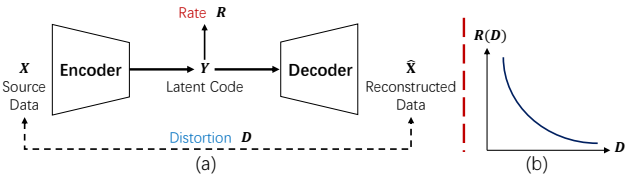

Figure 3. (a) Illustration of the data compression system and (b) Rate-Distortion trade-offs.

B. Background

B.1. Rate Distortion Theory

The Rate-Distortion theory [2, 3, 8] presents an analytical expression for the trade-off between the bit-rate and reconstruction quality of data compression. Fig. 3 demonstrates a typical system of lossy data compression, where *rate* indicates the bit length (size) of the latent code Y while *distortion* reflects the fidelity of reconstructed data. Rate-Distortion theory models the lower-bound of compression distortion with a given bit-rate. For source data X, $Y = Encoder(X)$ is the compressed code (latent code), and $\hat{X} = Decoder(Y)$ is the reconstructed data. The distortion can be measured by distortion function $D = \mathbb{E}_{P(X,\hat{X})}[\Delta(X,\hat{X})]$, where Δ is ℓ_1 loss in our case . Given

a maximum expected distortion D^*, the lower-bound for the bit-rate R is given by:

$$R(D^*) = \min_{D \leq D^*} \{I(X; \hat{X})\}, \tag{1}$$

where $I(X; \hat{X}) = H(X) - H(X \mid \hat{X})$ is the mutual information between source and reconstruction data. Shannon lower bound is defined as:

$$\begin{aligned} R(D^*) &= \min_{D \leq D^*} \{H(X) - H(X \mid \hat{X})\} \\ &= H(X) - \max_{D \leq D^*} \{H(X - \hat{X} \mid \hat{X})\} \\ &\geq H(X) - \max_{D \leq D^*} H(X - \hat{X}). \end{aligned} \tag{2}$$

This shows a larger bit-rate is needed for smaller distortion.

B.2. Information Bottleneck Theory.

Rate-Distortion theory determines the level of inevitable distortion D with a specific rate R. Information Bottleneck theory [9–11] further extends it without explicitly defining the distortion function:

$$R = \min\{I(X; Y) - \beta I(Y; \hat{X})\}. \tag{3}$$

This theory further conjectures that the training process of deep models consists of two stages. The network first fits on the training data, where $I(Y; \hat{X})$ increases and then forgets minor information, where $I(X; Y)$ decreases. This observation implies the essential role of forgetting in learning, and the deep model thus primarily learns common patterns of the training data for reconstruction. In contrast, infrequent patterns and image-specific details are typically forgettable trivialities for the trained models.

C. More Results

C.1. Additional Ablation Study

Visualization of ADA outputs. The visualization is given in Fig. 4. Thanks to the self-supervised learning scheme, ADA can be used to align missing details indicated in $\tilde{\Delta}$ (e.g., hat, hair in the last row). This process can lose some details (e.g., wrinkle, hair) in $\hat{\Delta}$ in case the misalignment is huge, which is one of the limitations of our method.

C.2. In-the-Wild Image Editing

We perform our inversion and editing framework on some Internet images. The results are shown in Fig. 5.

C.3. More Qualitative Comparison

We show more inversion and editing results on facial domain in Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 11, Fig. 10, Fig. 12, Fig. 13, Fig. 14, which involve editing on age, smile, eyes, lip, beard and pose. We also show inversion and editing results on car domain in Fig. 15, Fig. 16, Fig. 17 and Fig. 18, which involve editing on color and background. The proposed approach significantly outperforms baselines in terms of both inversion and editing performance. More comparisons with optimization-based are given in Fig.19. Note that the proposed approach is considerably faster ($\sim$ 1000$\times$) than these optimization methods. Fig. 20 also gives the comparison with a hybrid method.

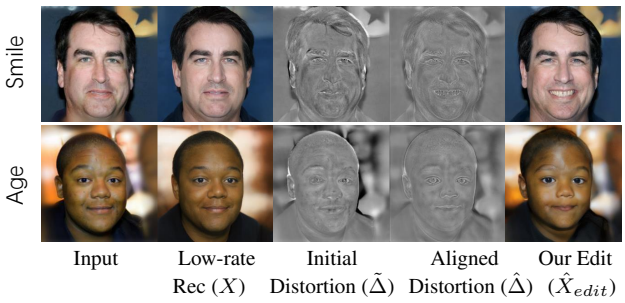

Figure 4. Visualization of ADA outputs for attribute editing.

C.4. Fine-grained Editing

We linearly increase the editing degree α for fine-grained attribute editing. As shown in Fig. 21, the editing results transit smoothly as α changes. See more results in the '*video-results.mp4*'.

C.5. Video Editing

More qualitative results on video inversion and editing are shown in Fig. 22. Following the pre-processing pipelines of CelebA-HQ and FFHQ, we first detect and crop the face regions from a video, and then use well-trained face image restoration models to improve the video quality (e.g., denoise, super-resolution) as our input. See more results in the '*video-results.mp4*'.

C.6. Limitation

As we adopt random transformation to simulate training pairs for ADA, the distortion map can be adaptively aligned for the editing of most attributes. However, for some challenging cases such as changing viewpoint, the misalignment can be significant and out-of-range. This thus leads to unsatisfactory editing in Fig. 23.
One possible remedy for this case is to explicitly pre-align the distortion map via landmarks or descriptors detection before ADA. For example, we can use facial landmarks detection approach to obtain landmarks that build sparse correspondence between two misaligned images. We can then interpolate the X-coordinates and Y-coordinates of sparse landmarks by Delaunay triangulation interpolation, as shown in Fig. 24. The interpolated coordinate maps establish dense correspondences and can be used to warp the distortion map to produce a coarsely aligned distortion,

Figure 5. Inversion and editing results on some Internet images.

which is further aligned by the ADA module for consultation.

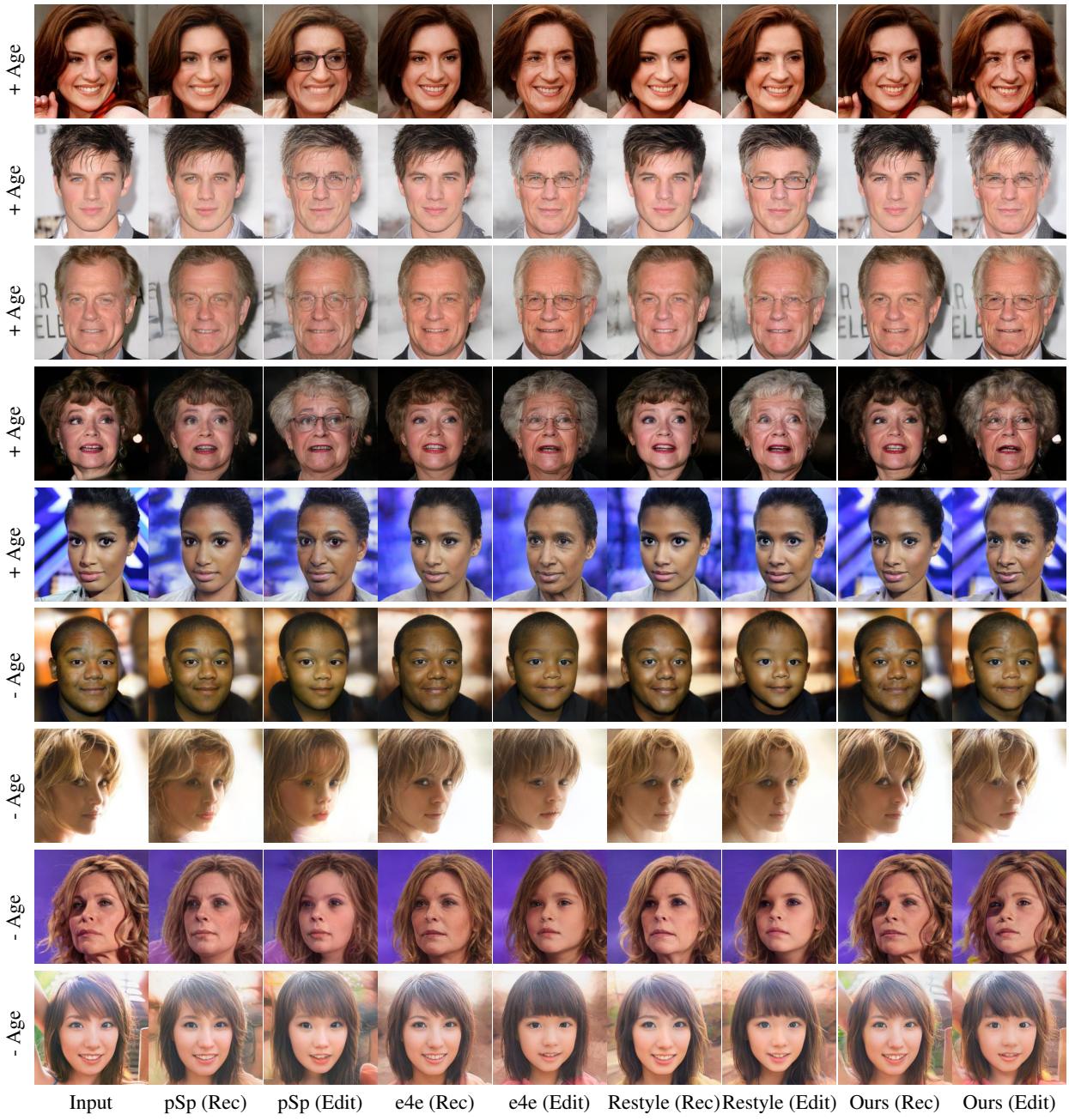

Figure 6. Visual comparisons on Face editing. (Age)

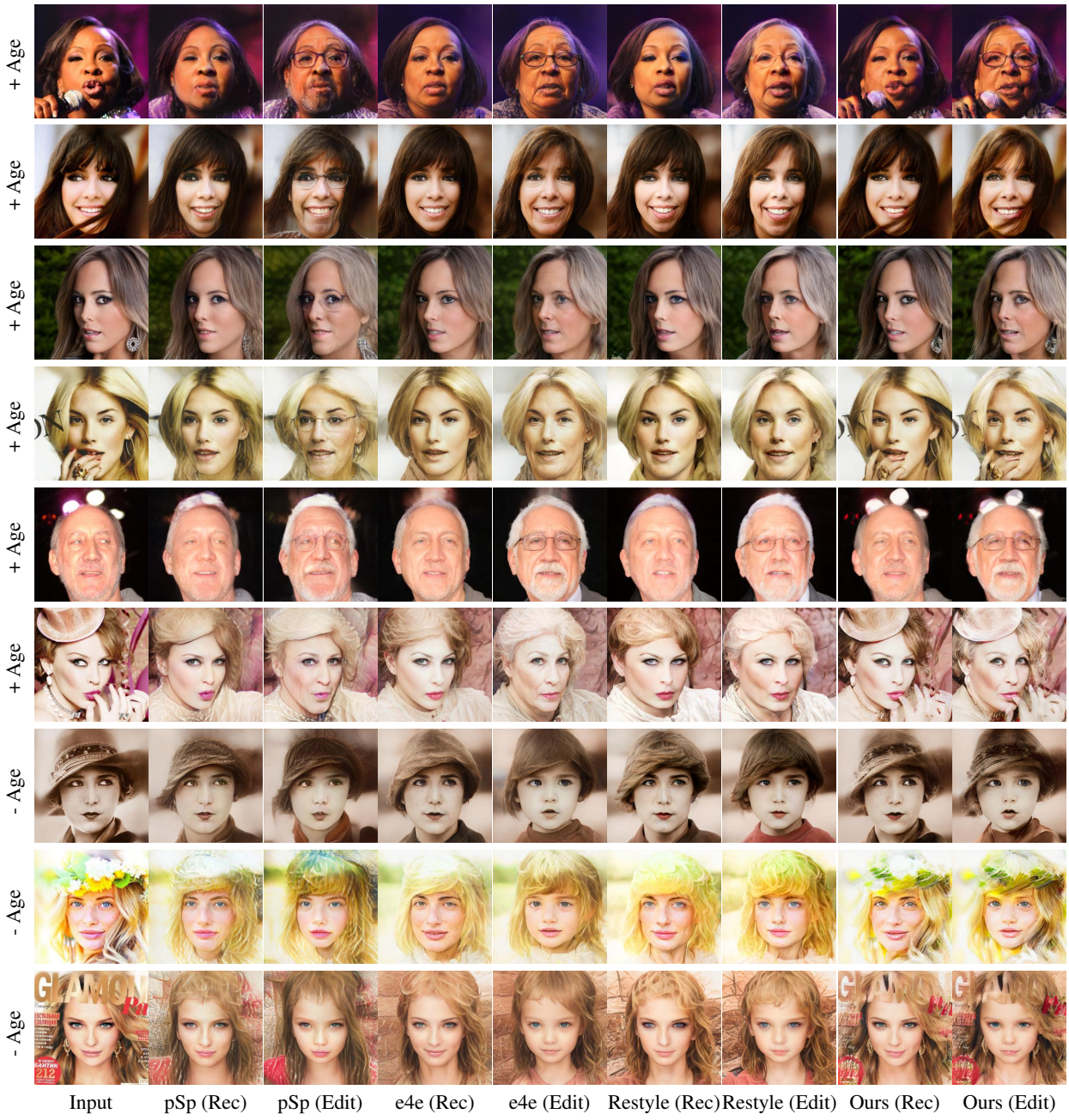

Figure 7. Visual comparisons on Face editing. (Age)

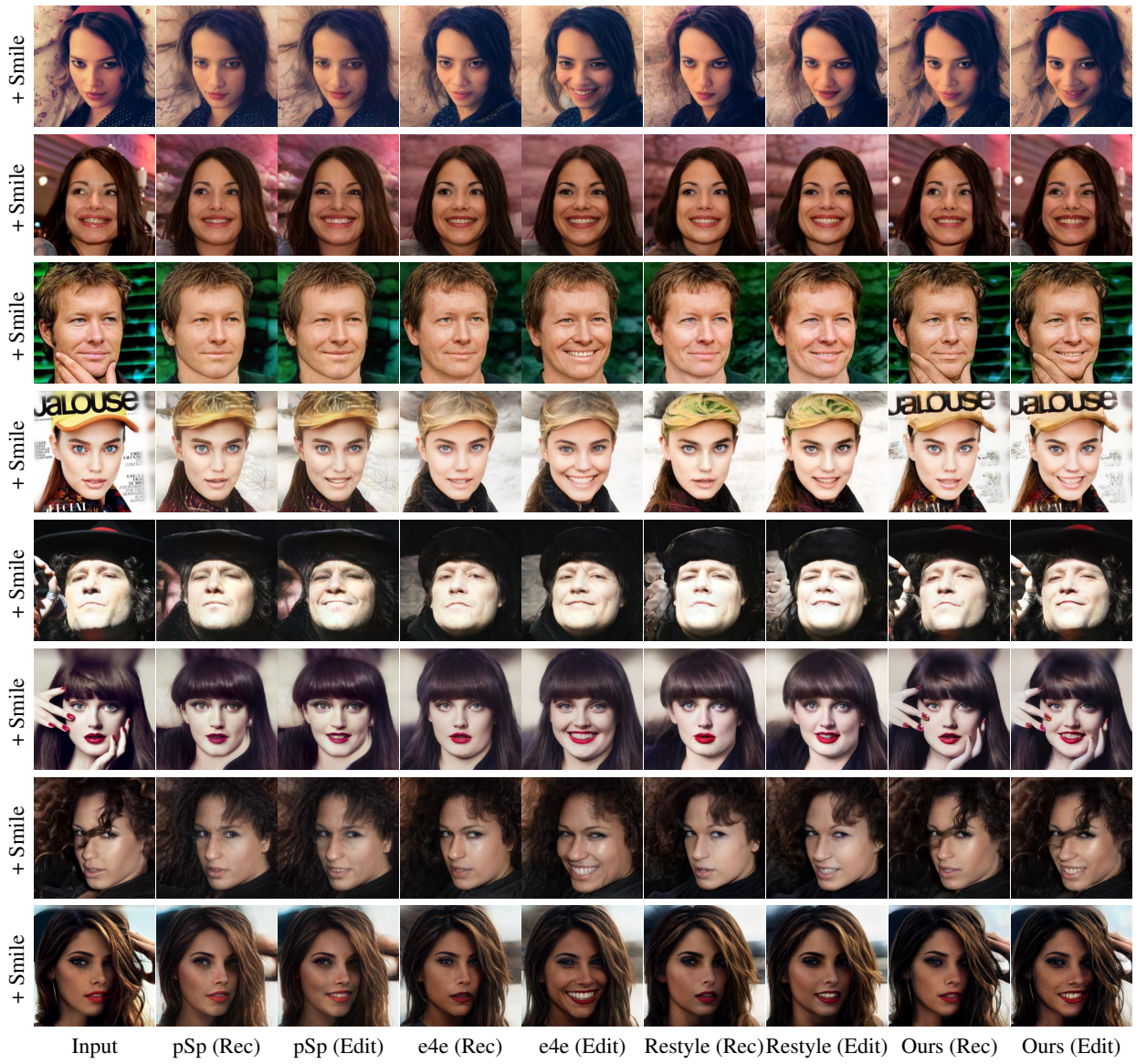

Figure 8. Visual comparisons on Face editing. (Smile)

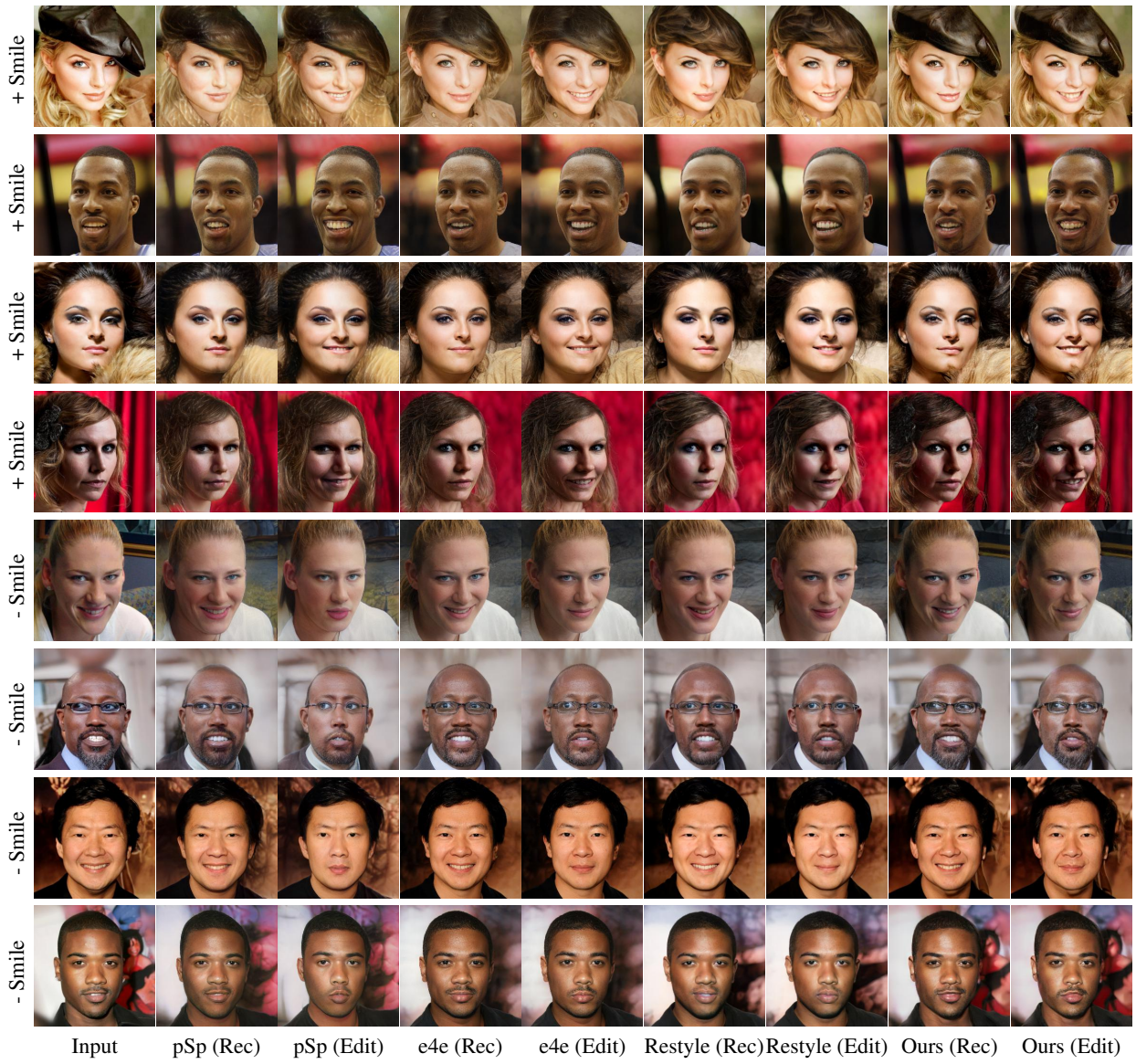

Figure 9. Visual comparisons on Face editing. (Smile)

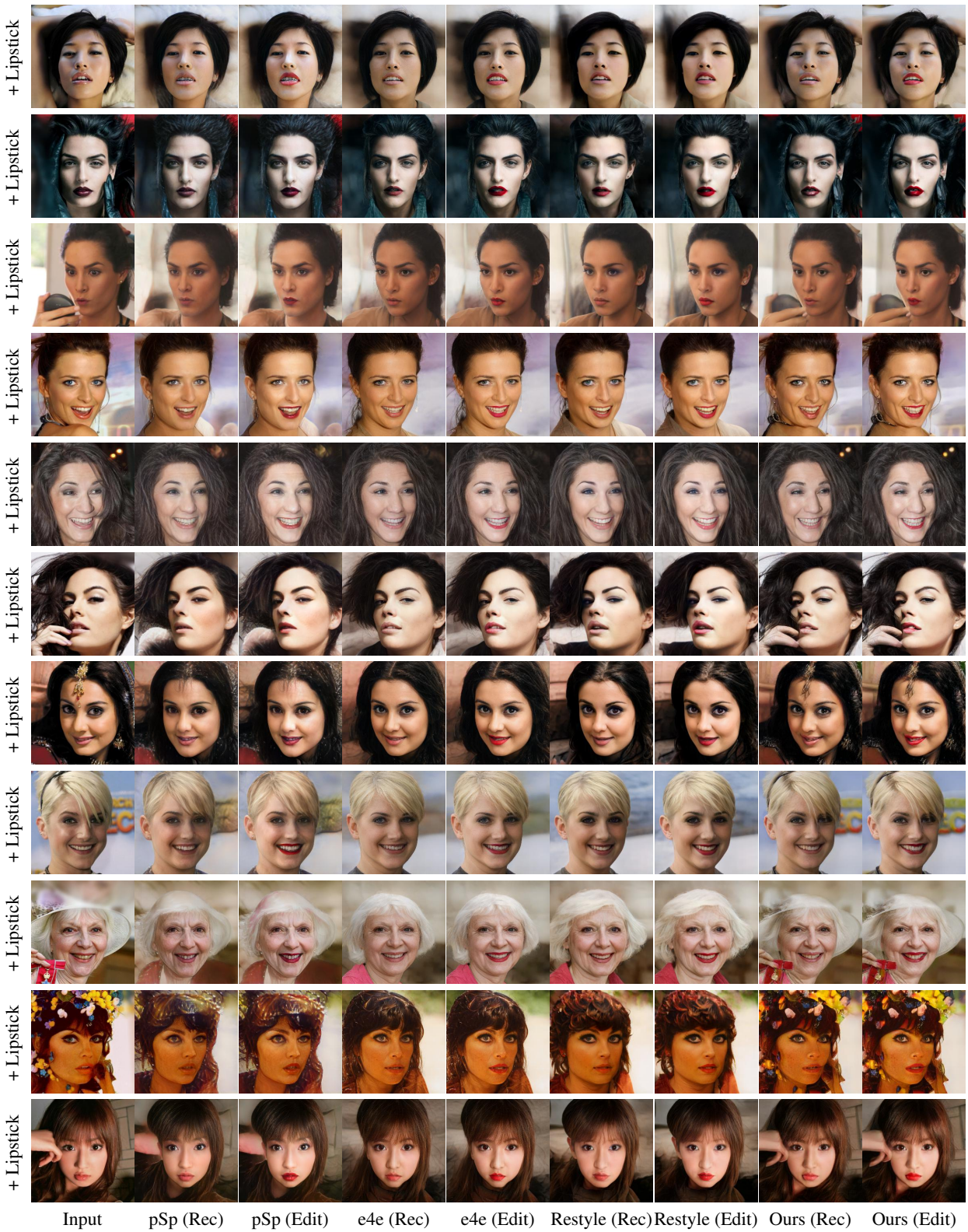

Figure 10. Visual comparisons on Face editing. (Lipstick)

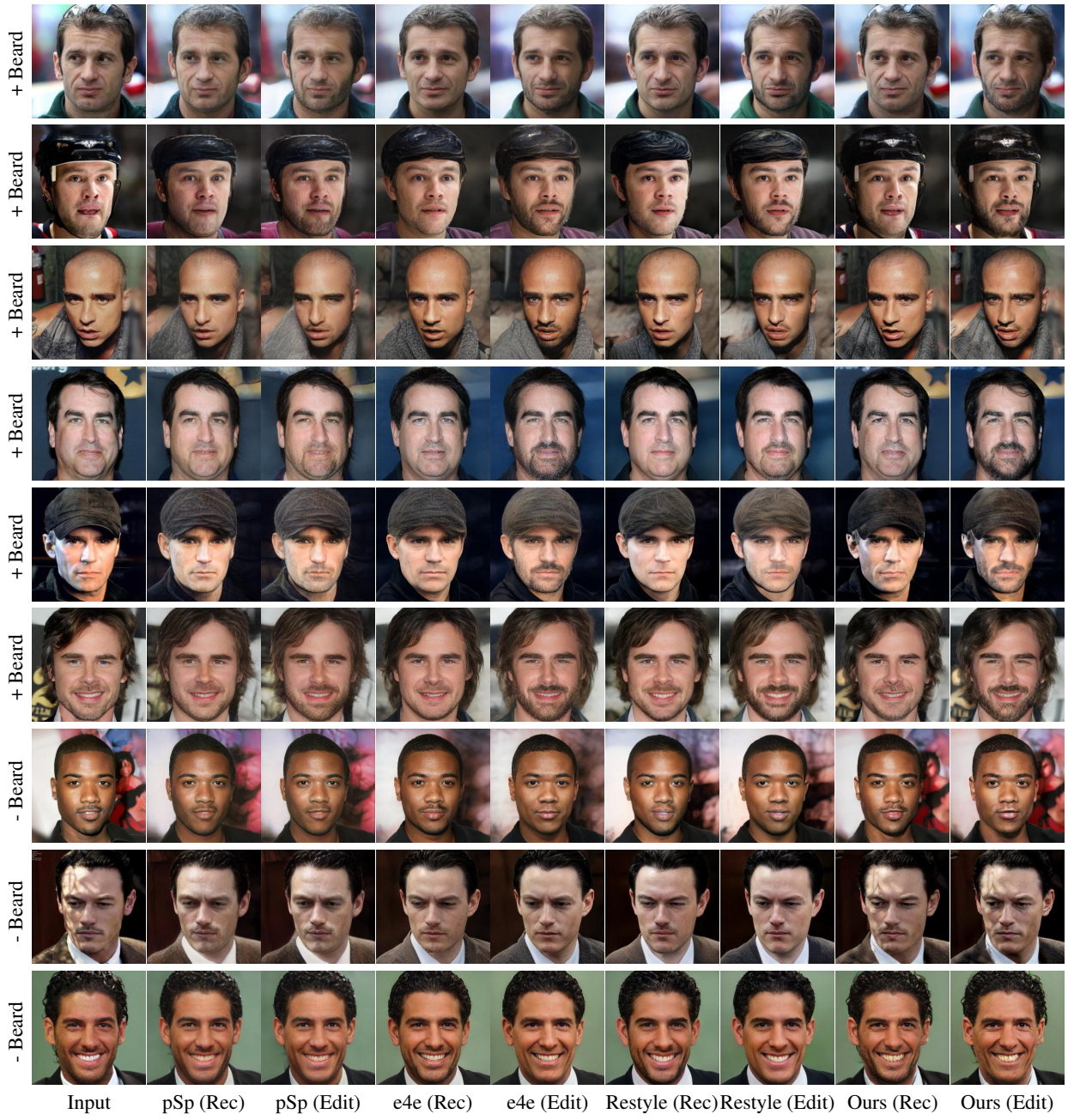

Figure 11. Visual comparisons on Face editing. (Beard)

Figure 12. Visual comparisons on Face editing. (Pose)

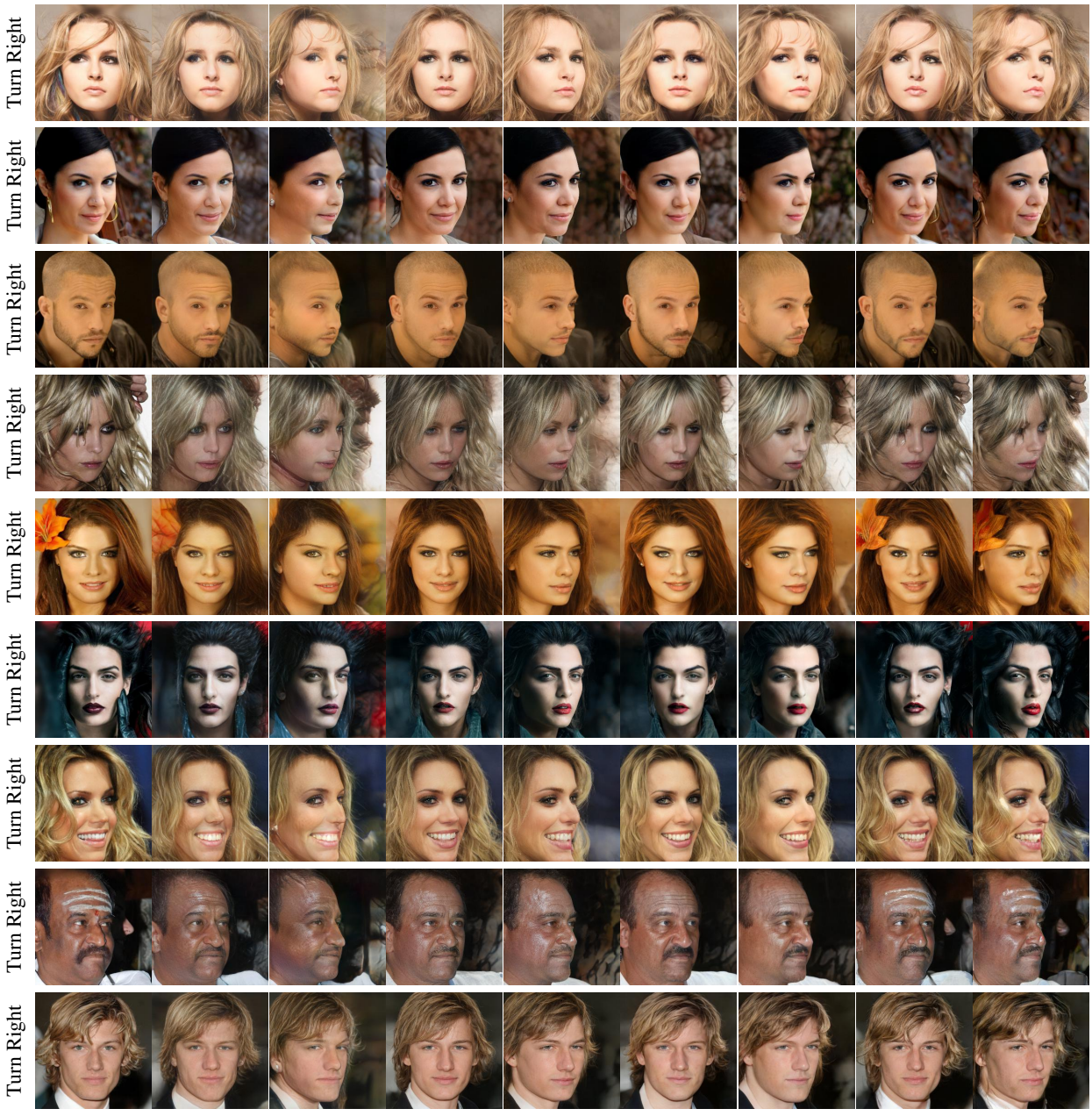

Figure 13. Visual comparisons on Face editing. (Pose)

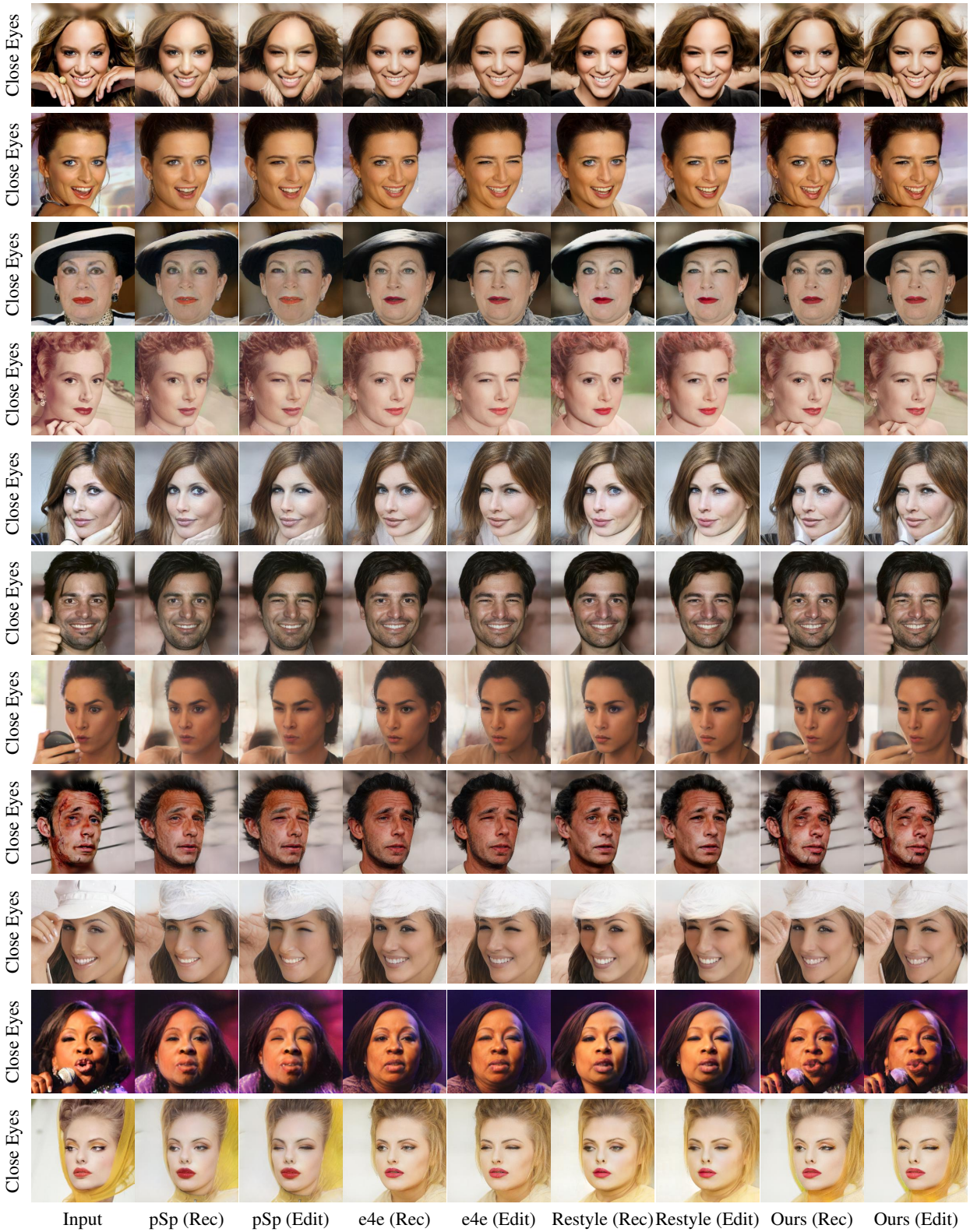

Figure 14. Visual comparisons on Face editing. (Eyes)

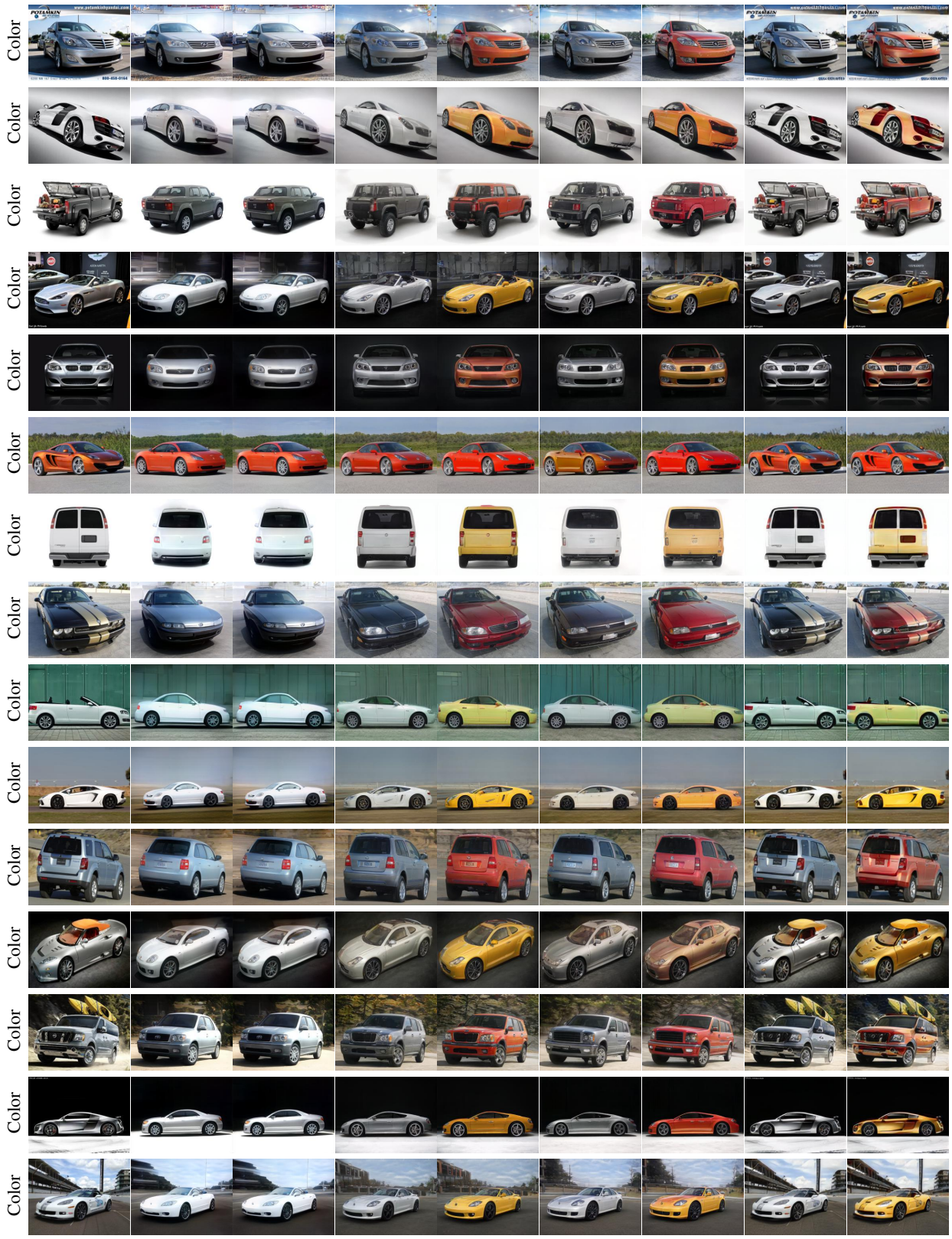

Figure 15. Visual comparisons on Car editing. (Color)

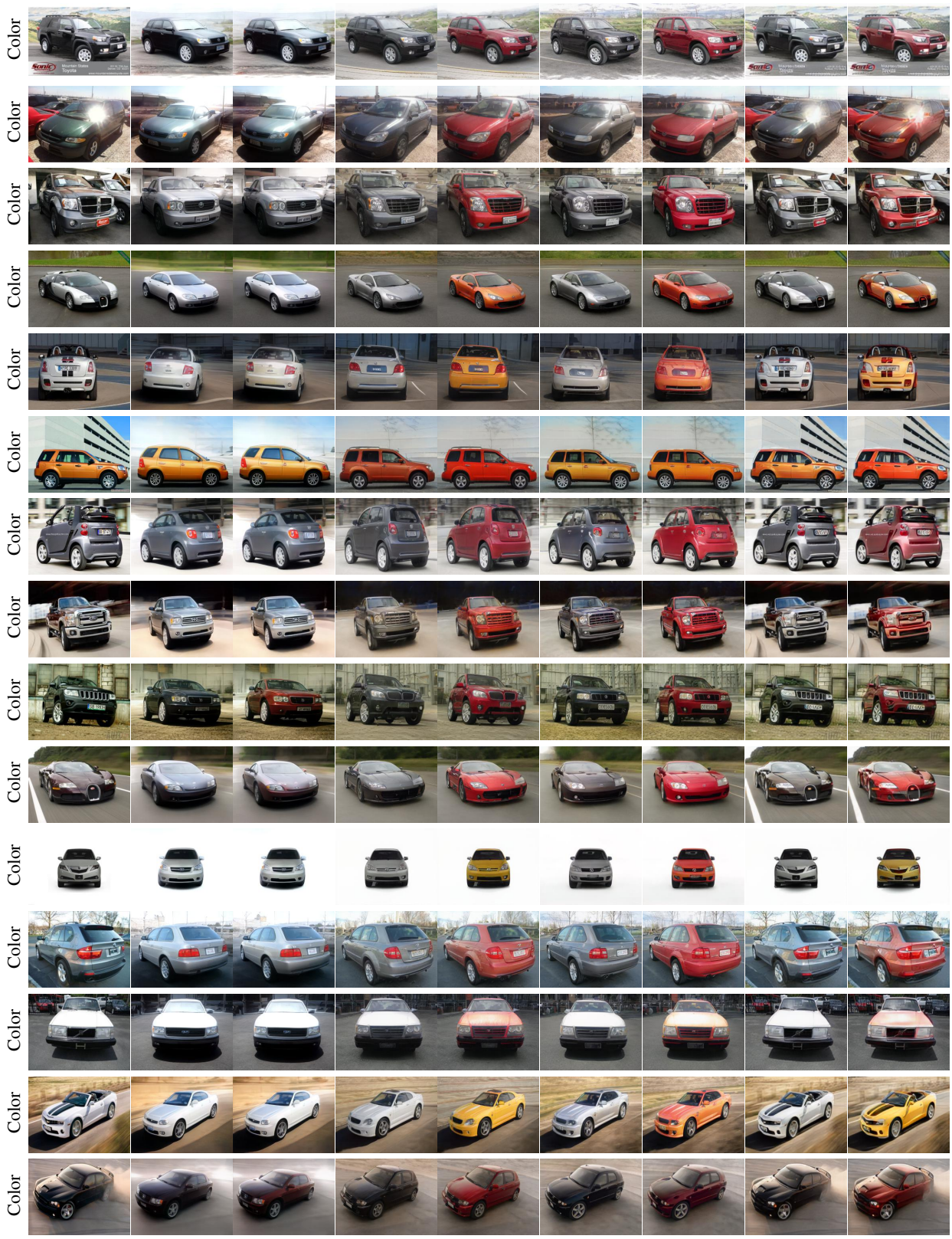

Figure 16. Visual comparisons on Car editing. (Color)

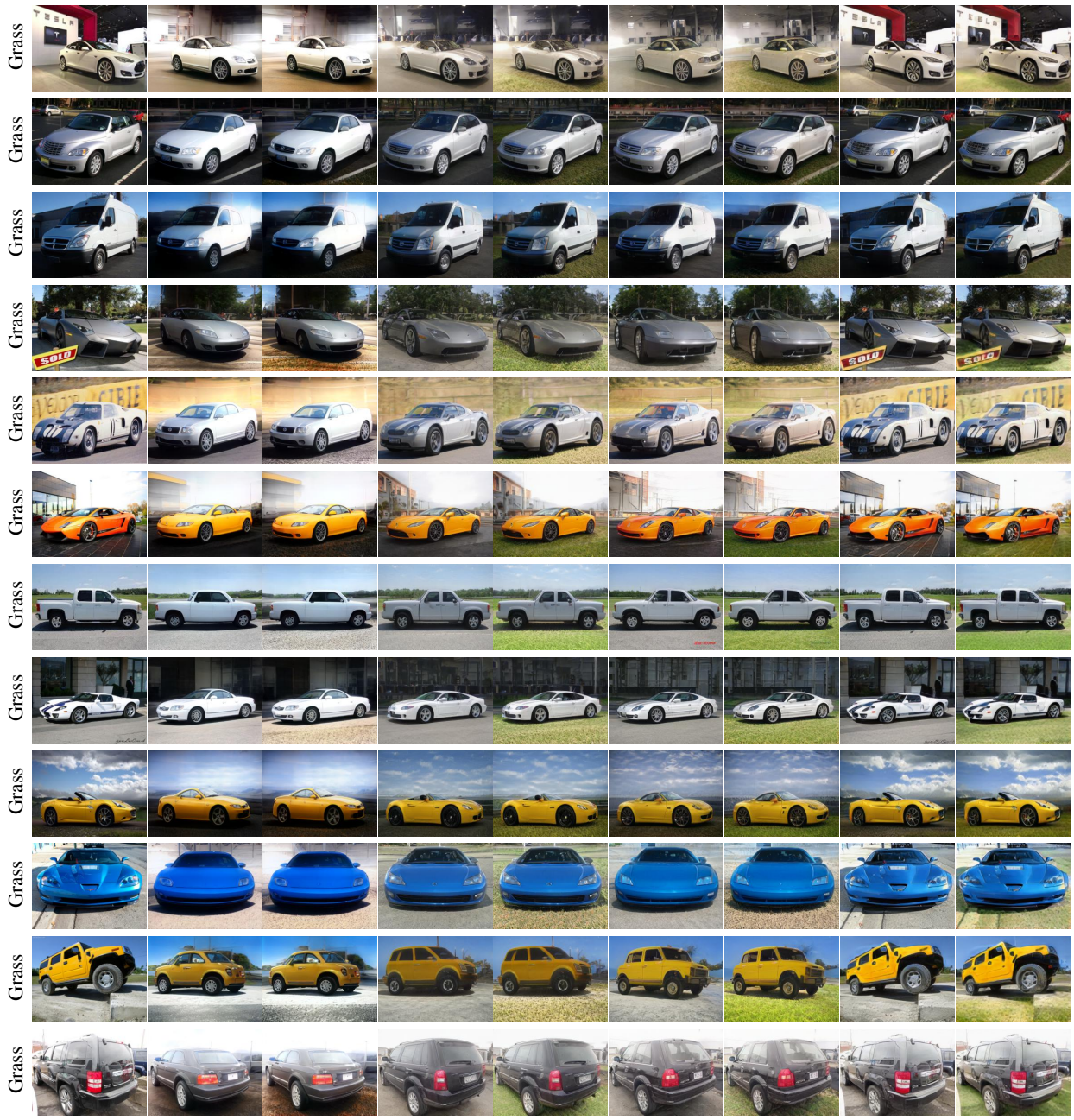

Figure 17. Visual comparisons on Car editing. (Grass)

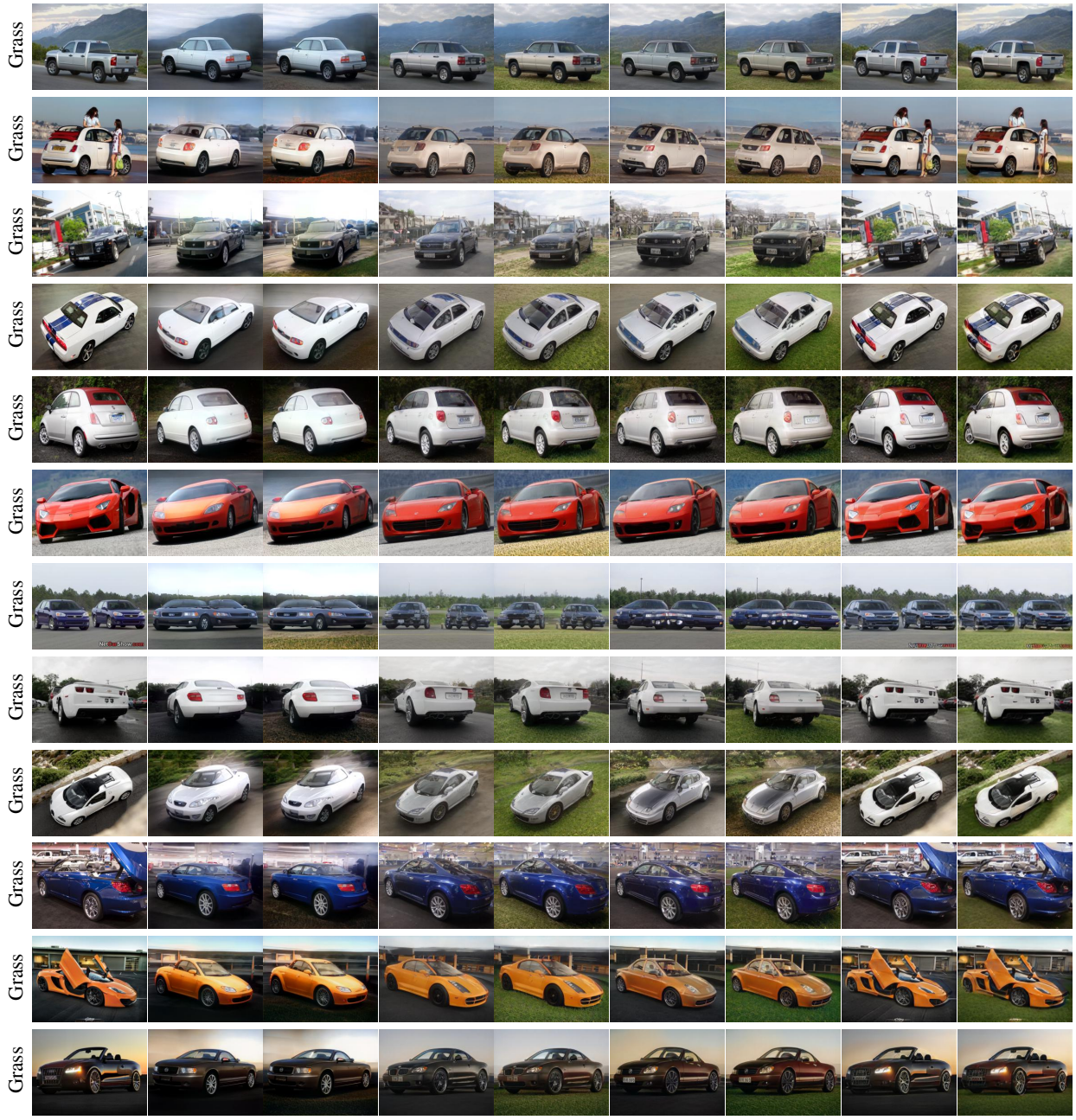

Figure 18. Visual comparisons on Car editing. (Grass)

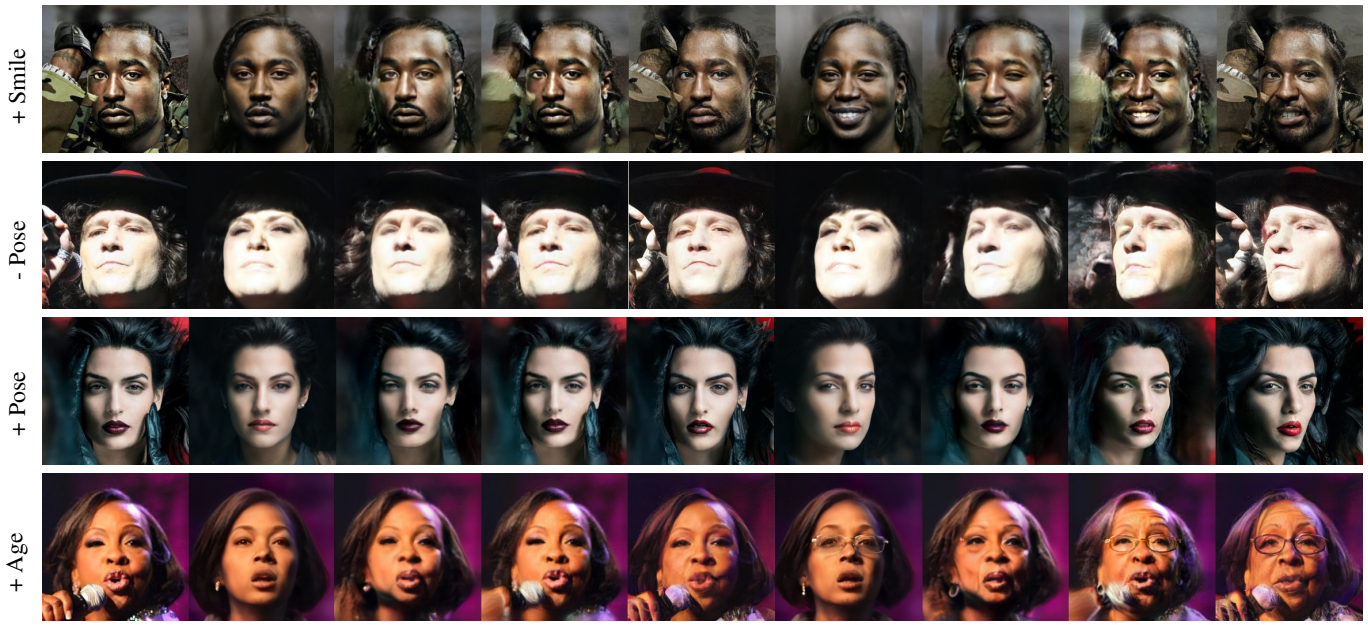

Figure 19. Visual comparison with optimization-based methods.

Figure 20. Comparison with a hybrid method (In-domain GAN [14]).

Figure 21. Results of fine-grained attribute editing. We linearly interpolate the editing degree α and perform editing. See more results in the attached '*video-results.mp4*'.

Figure 22. Results of inversion and editing (+ age and + smile) on videos. See more results in '*video-results.mp4*'.

Figure 23. Failure Cases. Our scheme sometimes suffers artifacts when sharply manipulating image viewpoints.

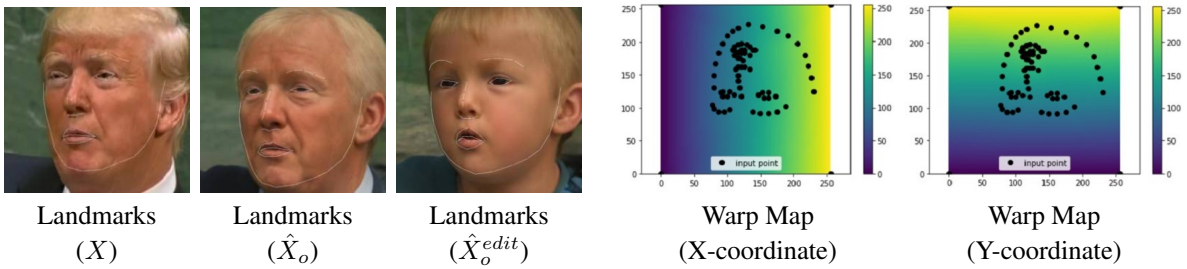

Figure 24. Example of facial landmarks detection and coordinate interpolation.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 17

[2] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning (ICML)*, 2019. 1

[3] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 1

[4] 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 (CVPR)*, 2019. 17

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations (ICLR)*, 2015. 1

[7] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021. 17

[8] Claude E Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959. 1

[9] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. 2

[10] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *Proceedings of Annual Allerton Conference on Communication, Control and Computing*, 1999. 2

[11] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015. 2

[12] 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. 1

[13] Michael R Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back. *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 1

[14] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European Conference on Computer Vision (ECCV)*, 2020. 17