Supplementary Material of CLCAE

Hongyu Liu[1] Yibing Song[2*] Qifeng Chen[1*]
[1]Hong Kong University of Science and Technology [2]AI[3] Institute, Fudan University
hliudq@cse.ust.hk yibingsong.cv@gmail.com

In this supplementary material, we first describe the limitation of our method. Then, we present more analysis about our solid foundation latent code w. Meanwhile, we show more visual comparisons of the ClebA-HQ [7] and car datasets [13]. Finally, we demonstrate our method can achieve good performance in the horse dataset [13] in visually.

1. Limitation

Our method has good performance in both qualitative and quantitative, but it still has some limitations. Our method cannot reconstruct the jewelry well of some corner cases, and there are some artifacts during the editing process. We can replace the CNN with a more powerful network(i.e., Vision Transformer [3, 6]) to try to solve these problems.

2. More Analysis

To further prove that our method can predict robust latent code w. We set our w as the initialization of PTI [9] to make comparisons. As shown in Fig. 1, the (a) is the original initialization results with w in PTI, and PTI finds this w with the optimization method [5]. The (b) is the reconstruction results with our w, and the (b) outperformance than (a) in both identity and detail preservation which verifies the effectiveness of our method. The (c) is the original final prediction of PTI which sets the optimization w as the initialization, and we replace the optimization w with our w to get (d). By comparing (c) and (d), we can find a robust w that can improve the performance of PTI. Meanwhile, since the w in (d) is predicted with our encoder, we can speed PTI up to 134s for a single image, which is almost half of the time-consuming of the original PTI. Moreover, we provide more visual results of ablation study 2.

3. More visual comparisons

$\mathcal{W}^+$ space. We show more visual comparisons between $\mathcal{W}^+$ space methods (e4e [10], pSp [8], restyle$_{\text{pSp}}$ [1], restyle$_{\text{e4e}}$ [1] and StyleTransformer (ST) [4]) and our method in Fig. 3 and Fig. 4. Except for the e4e and our method, the other methods seem to have an overfitting phenomenon (i.e., the wrong white hair in the (c),(d), and (e) of the second person in Fig. 3) as discussed in our main paper. Meanwhile, our method has better reconstruction and editing performance simultaneously than other baselines (i.e., the ”Age” and ”Smile” editing results in Fig. 3 and the ”Viewpoint” editing results in Fig. 4).

$\mathcal{F}$ space. Fig. 5 and Fig. 6 shows more our comparisons to PTI [9], Hyper [2], HFGI [11], and FS [12] in the $\mathcal{F}$ space. Our method can produce the image with better quality in both reconstruction and editing than other baselines (i.e., the ”Pose” editing results in Fig. 5 and the ”Grass” editing results in Fig. 6).

Moreover, we show more visual comparisons in Fig. 7.

4. More visual results

In addition to the face and car datasets, we also show more visual results on horse dataset [13] in Fig 8. We show the reconstruction results with our w, w^+ and w^+, f in (a) , (b) and (c) respectively. These visual results show that our solid foundation latent code w method can produce good-quality reconstruction images, and our w^+ and f can further generate high-fidelity results with the solid w.

*Y. Song and Q. Chen are corresponding authors.

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021.

[2] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Trans-

Figure 1. Analysis of latent code w. We replace the initialization of PTI with our w as shown in (d). The original PTI's result is (c). We can find that our solid latent code w can help the PTI perform better. Meanwhile, we illustrate the reconstruction results with optimization w and our w in (a) and (b), respectively.

Figure 2. Qualitative ablation

formers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Xueqi Hu, Qiusheng Huang, Zhengyi Shi, Siyuan Li, Changxin Gao, Li Sun, and Qingli Li. Style transformer for image inversion and editing. *arXiv preprint arXiv:2203.07932*, 2022.

[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*, pages 8110–8119, 2020.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild, 2015.

Figure 3. More visual comparisons on ClebA-HQ [7] dataset for $\mathcal{W}^+$ space methods. Our method performance better in both reconstruction and editing. ↓ means a reduction of the manipulation attribute. ↑ means an increment of the manipulation attribute.

[8] 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*, 2021.

[9] 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.

[10] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation, 2021.

[11] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[12] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A style-based gan encoder for high fidelity reconstruction of images and videos. *European conference on computer vision*, 2022.

[13] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2016.

Figure 4. More visual comparisons on car dataset [13] for $\mathcal{W}^+$ space methods. Our method performance better in both reconstruction and editing.

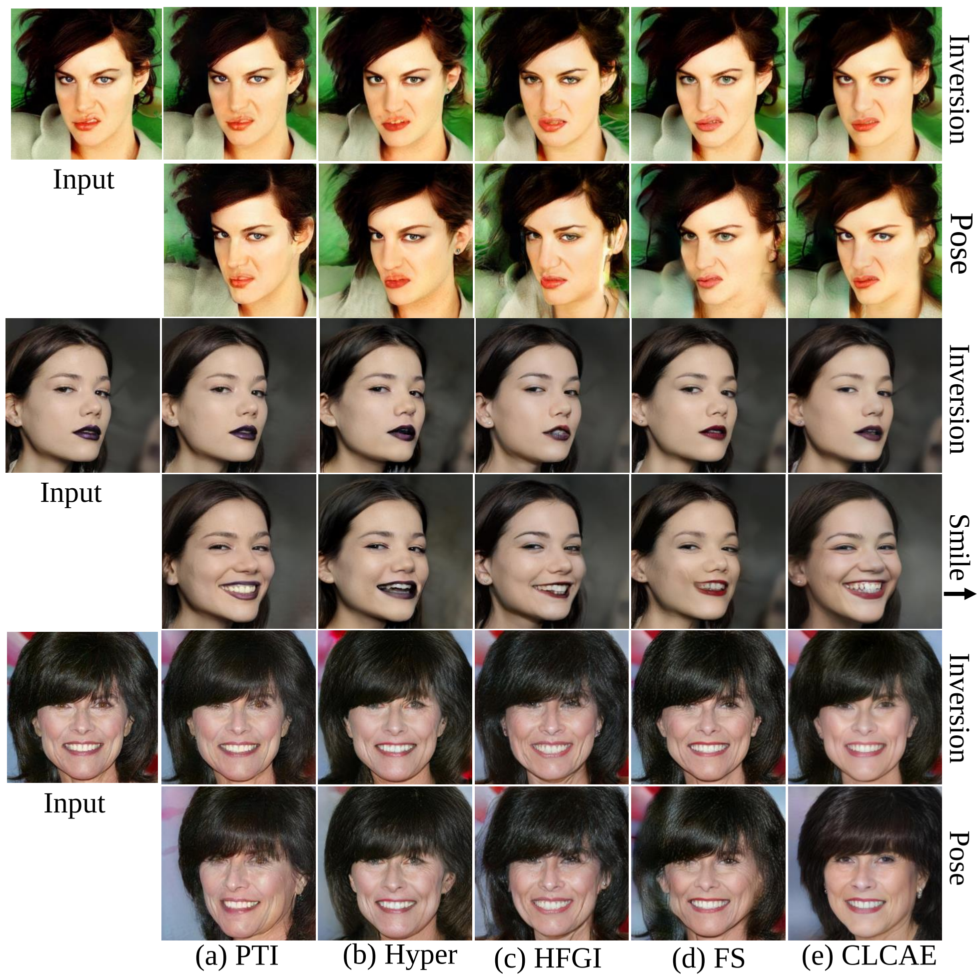

Figure 5. More visual comparisons on ClebA-HQ [7] dataset for $\mathcal{F}$ space methods. Our method performance better in both reconstruction and editing. ↑ means an increment of the manipulation attribute.

Figure 6. More visual comparisons on car dataset [13] for $\mathcal{F}$ space methods. Our method performance better in both reconstruction and editing.

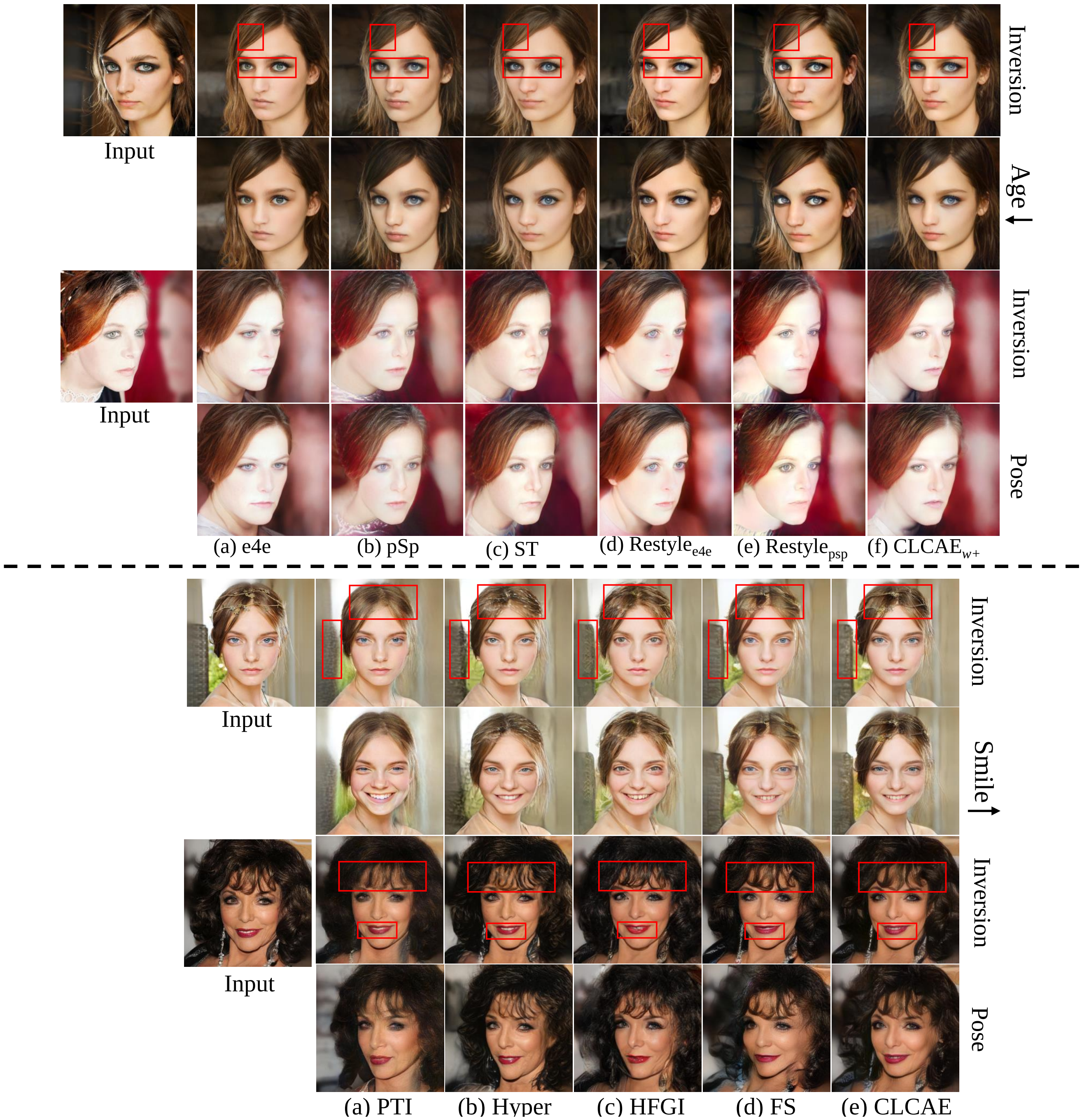

Figure 7. More visual comparisons.

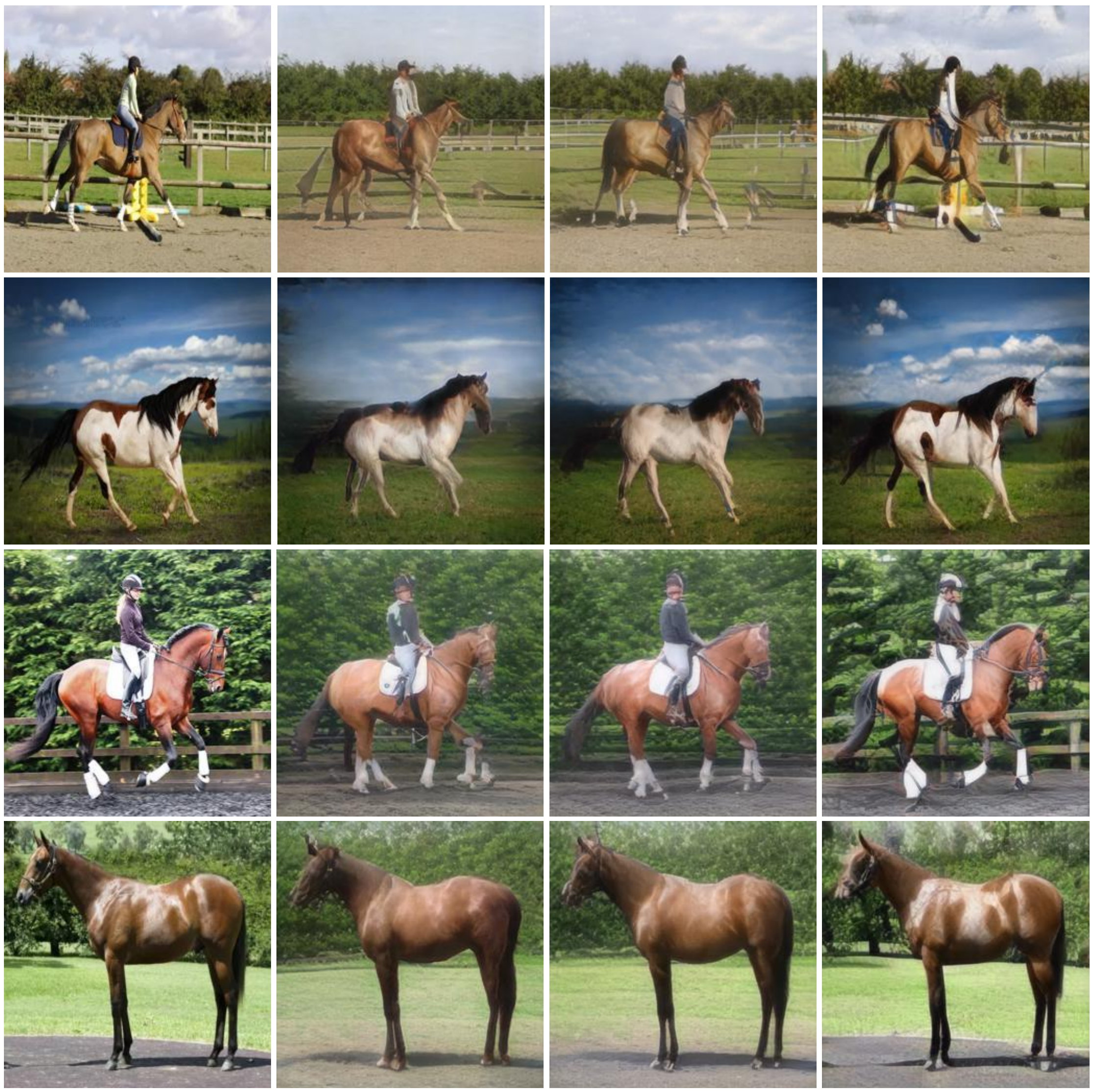

Figure 8. More visual results on horse dataset [13]. Good results can demonstrate the robustness of our method.