In-Domain GAN Inversion for Real Image Editing Supplementary Material

Jiapeng Zhu*$^{1[0000-0001-9198-0304]}$, Yujun Shen*$^{1[0000-0003-3801-6705]}$,
Deli Zhao$^{2[0000-0002-8838-578X]}$, and Bolei Zhou$^{1[0000-0003-4030-0684]}$

[1] The Chinese University of Hong Kong
{jpzhu, sy116, bzhou}@ie.cuhk.edu.hk
[2] Xiaomi AI Lab
zhaodeli@gmail.com

1 Overview

This supplementary material is organized as follows: In Sec.2, we show image reconstruction results using the model trained on LSUN bedroom dataset [5]. In Sec.3 and Sec.4, we show more results of image interpolation and image manipulation to verify that *in-domain* GAN inversion can recover the target images from both the pixel level and the semantic level. In Sec.5, we show some style mixing results. In Sec.6, we make detailed analysis on the semantic diffusion achieved by our *in-domain* inversion.

2 Image Reconstruction

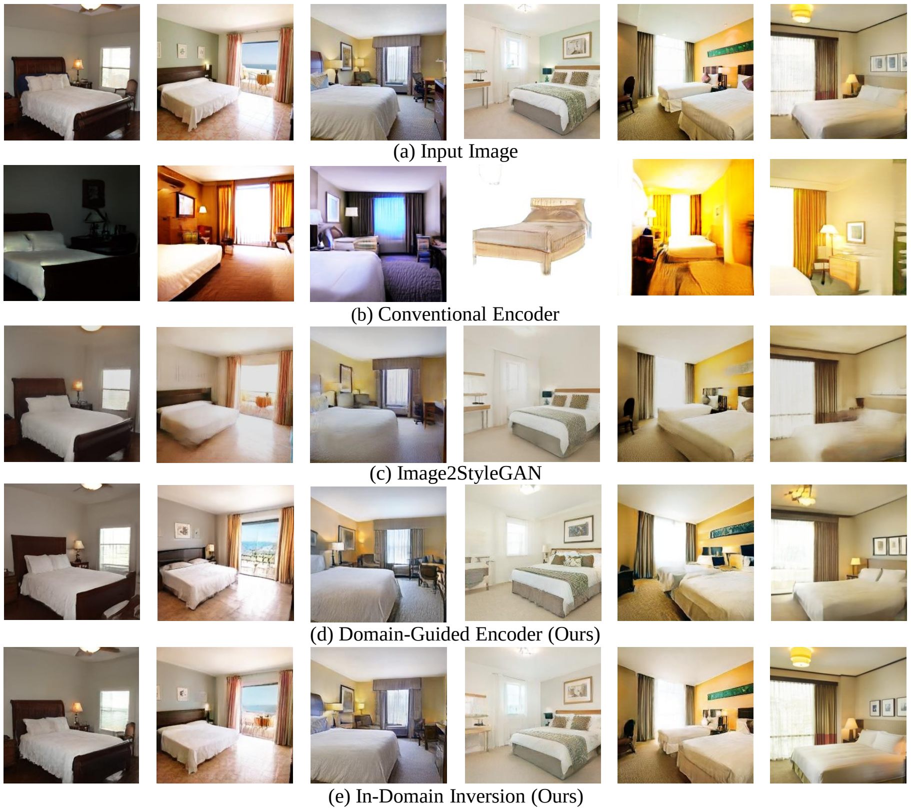
Image reconstruction is one of the most important metrics to evaluate a GAN inversion method. Besides human faces and towers (outdoor scene) shown in the main paper, we also do experiments on bedrooms (indoor scene) and compare with existing inversion methods. The comparison results are shown in Fig.1. We can tell that our proposed *domain-guided* encoder produces much better reconstructions than the conventional encoder [6]. The further *domain-regularized* optimization also surpasses the start-of-the-art optimization-based inversion method, Image2StyleGAN [2], with higher reconstruction quality.

3 Image Interpolation

Different from previous GAN inversion approaches that mainly focus on the image reconstruction from the pixel level, we propose to align the inverted code with the semantic knowledge learn by GAN models, *i.e.*, *in-domain*.

In this section, we use image interpolation to evaluate whether the inverted codes are semantically meaningful. Fig.2, Fig.3, and Fig.4 show the comparison

* denotes equal contribution.



(a) Input Image

(b) Conventional Encoder

(c) Image2StyleGAN

(d) Domain-Guided Encoder (Ours)

(e) In-Domain Inversion (Ours)

Fig. 1. Qualitative comparison on bedroom reconstruction with different GAN inversion methods. (a) Input image. (b) Conventional encoder [6]. (c) Image2StyleGAN [2]. (d) Our proposed *domain-guided* encoder. (e) Our proposed *in-domain* inversion.

results between Image2StyleGAN [2] and our *in-domain* inversion on faces, towers (outdoor scene), and bedrooms (indoor scene) respectively. We observe that the interpolations from Image2StyleGAN show unsatisfying artifacts and blurs, especially when the source and target images are with large discrepancy (*e.g.*, the first and last sample in Fig.2). Meanwhile, some interpolations made by Image2StyleGAN are not semantically meaningful (*e.g.*, interpolated images are no longer a tower any more in the last sample in Fig.3). On the contrary, our method makes sure that all interpolated samples are still with high quality and explanatory semantics. We show more results in Fig.5 (face), Fig.6 (tower), and Fig.7 (bedroom). In Fig.5, we manage to interpolate male and female, faces with different poses, or even painting and real person. In Fig.6, we interpolate one type of tower to other types in a large diversity. Each individual interpolation is realistic enough for a "new type" of tower. In Fig.7, we can interpolate between bedrooms from different viewpoints. It is also noteworthy that windows and paintings on the wall can also be adequately interpolated using our method.

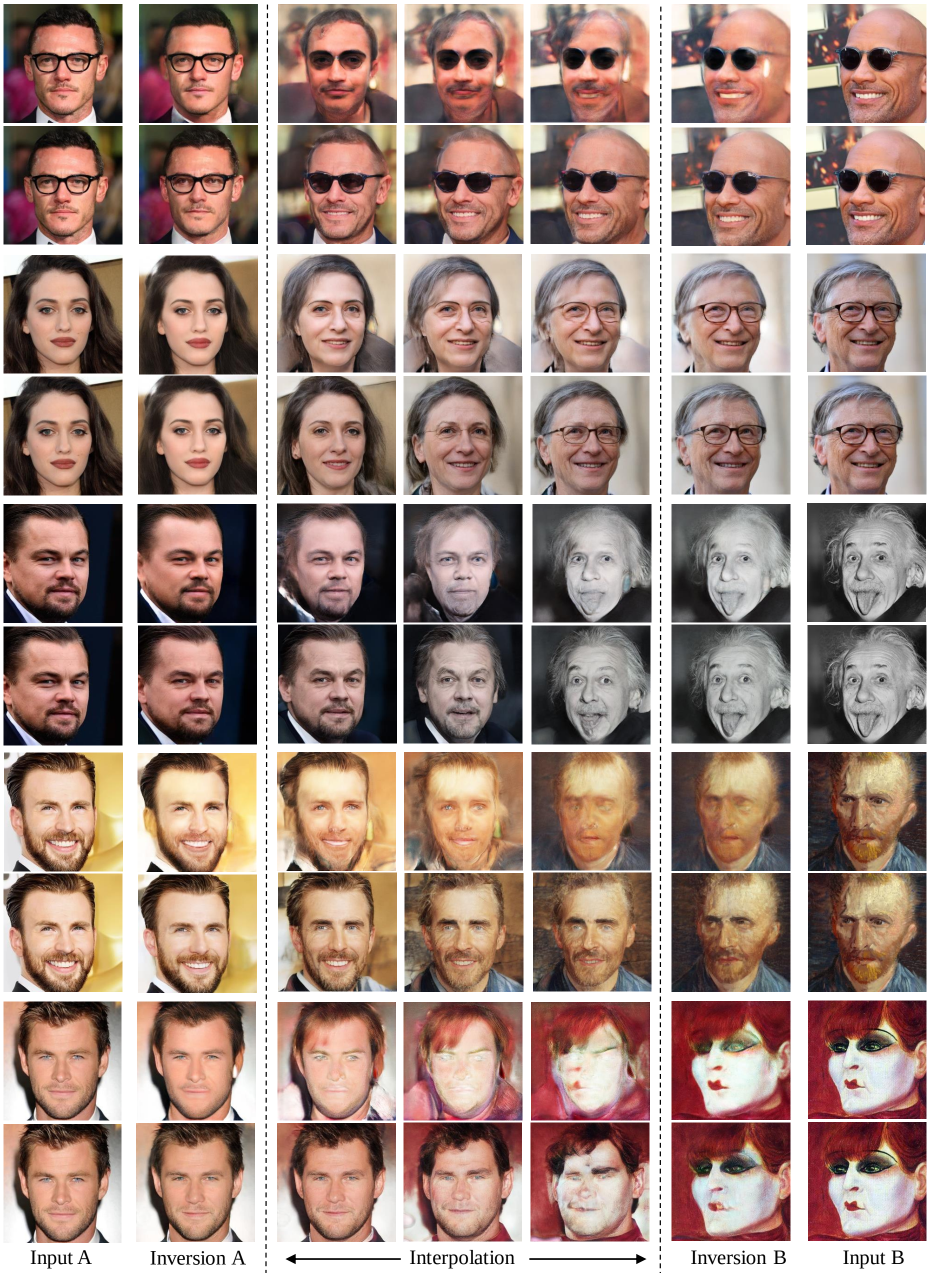


Fig. 2. Qualitative comparison on face interpolation between Image2StyleGAN [2] (odd rows) and our *in-domain* inversion (even rows). Zoom in for details.

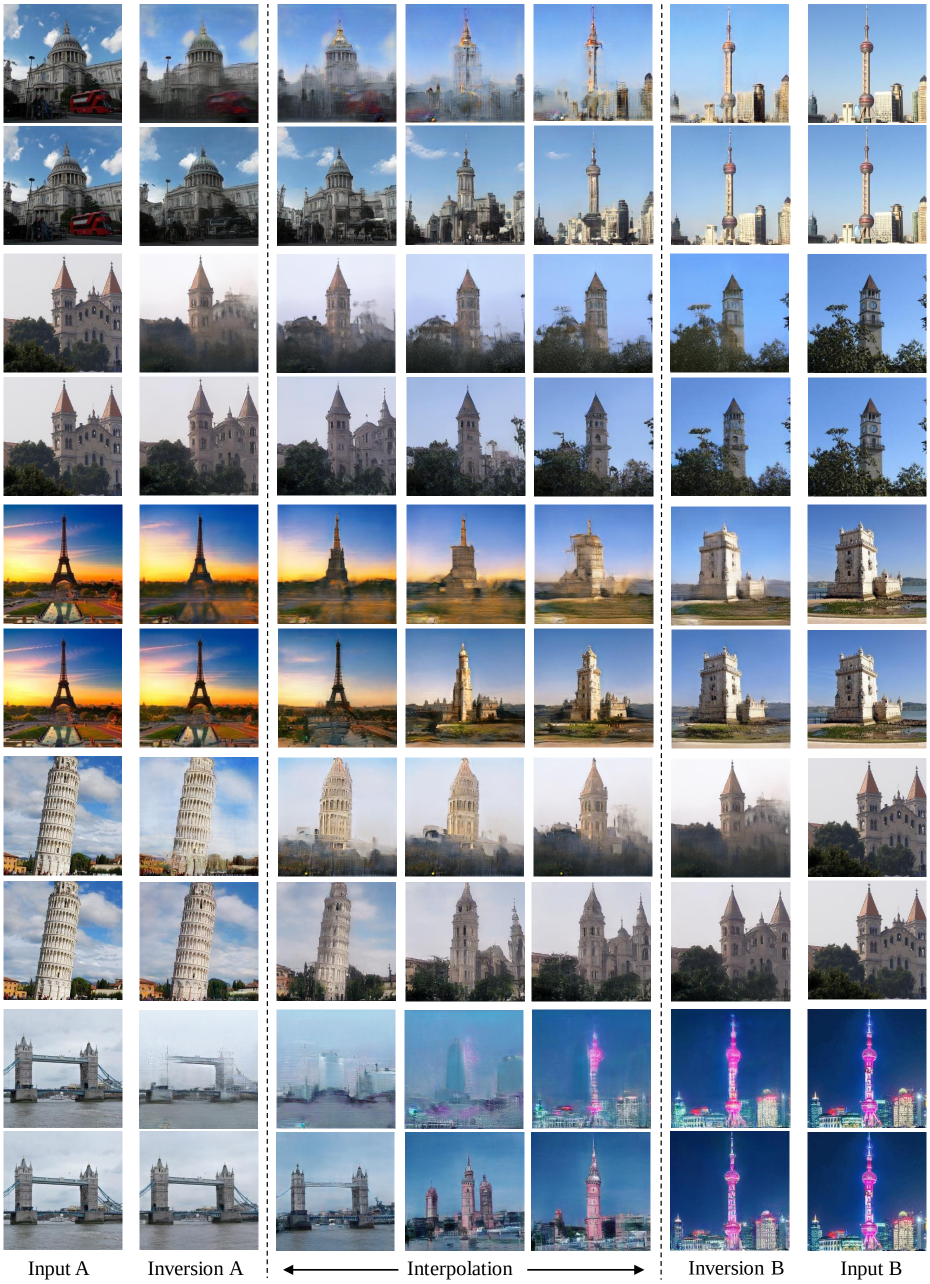


Fig. 3. Qualitative comparison on tower interpolation between Image2StyleGAN [2] (odd rows) and our *in-domain* inversion (even rows). Zoom in for details.

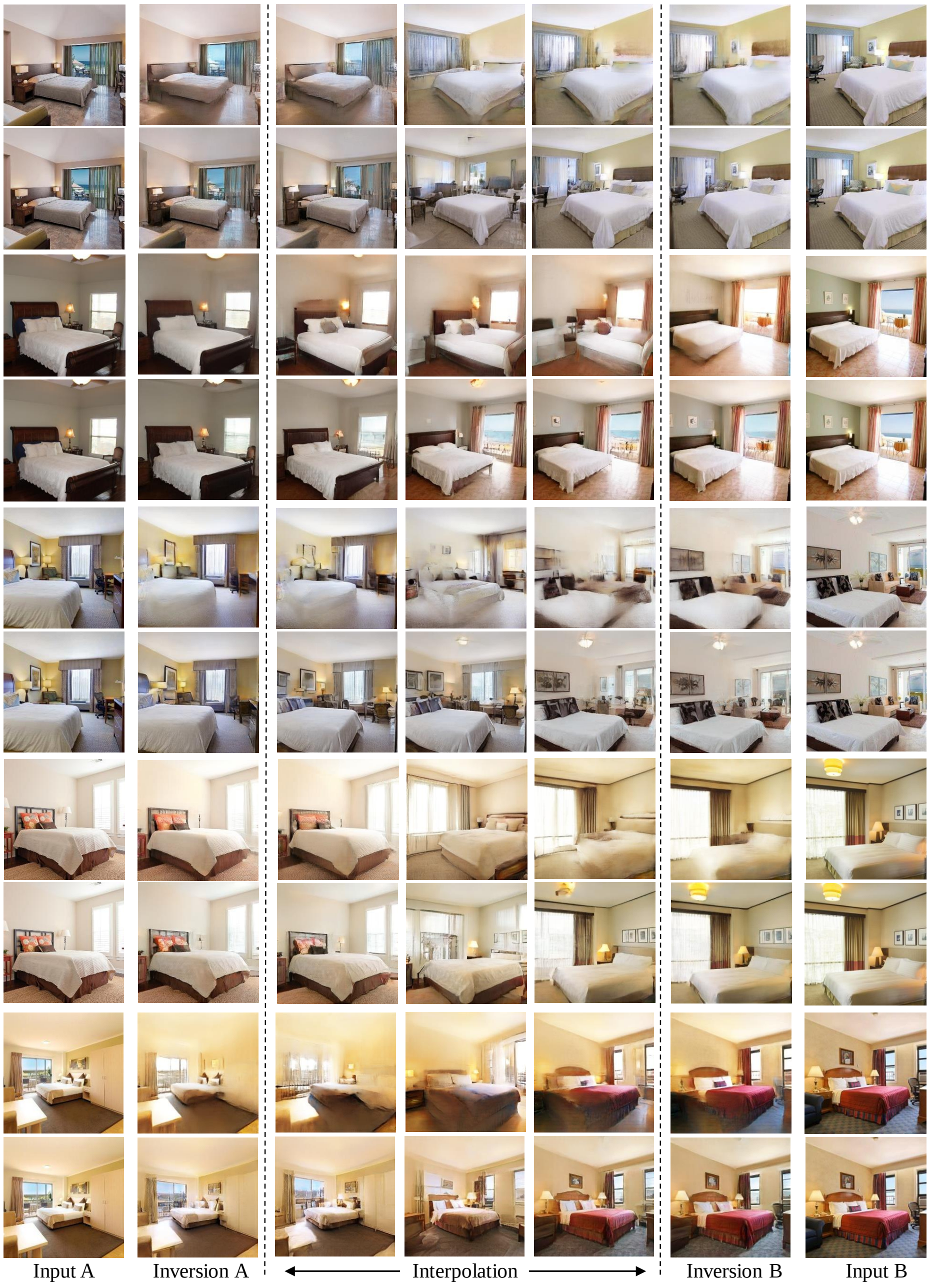


Fig. 4. Qualitative comparison on bedroom interpolation between Image2StyleGAN [2] (odd rows) and our *in-domain* inversion (even rows). Zoom in for details.

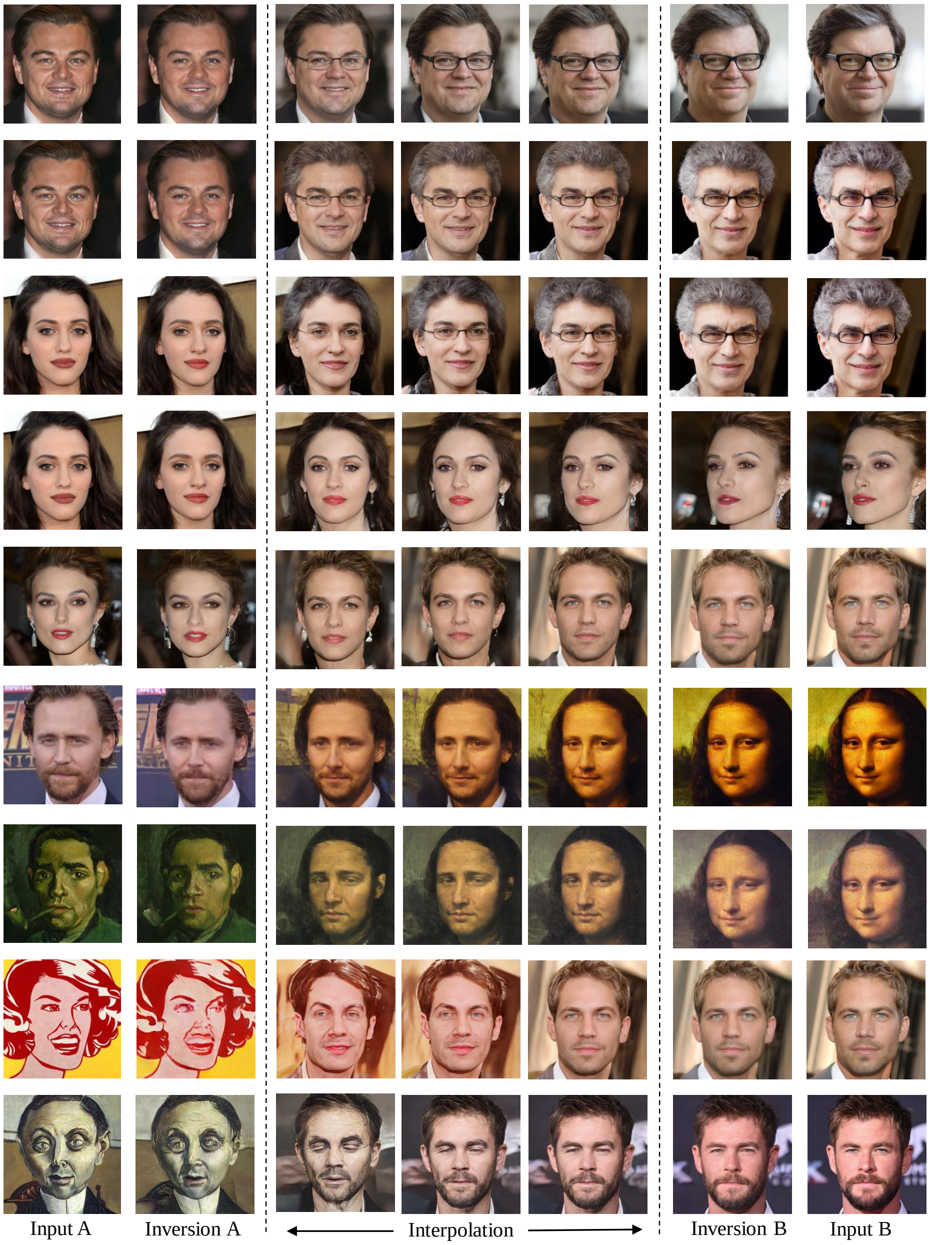


Fig. 5. Face interpolation results using our *in-domain* GAN inversion method. Zoom in for details.

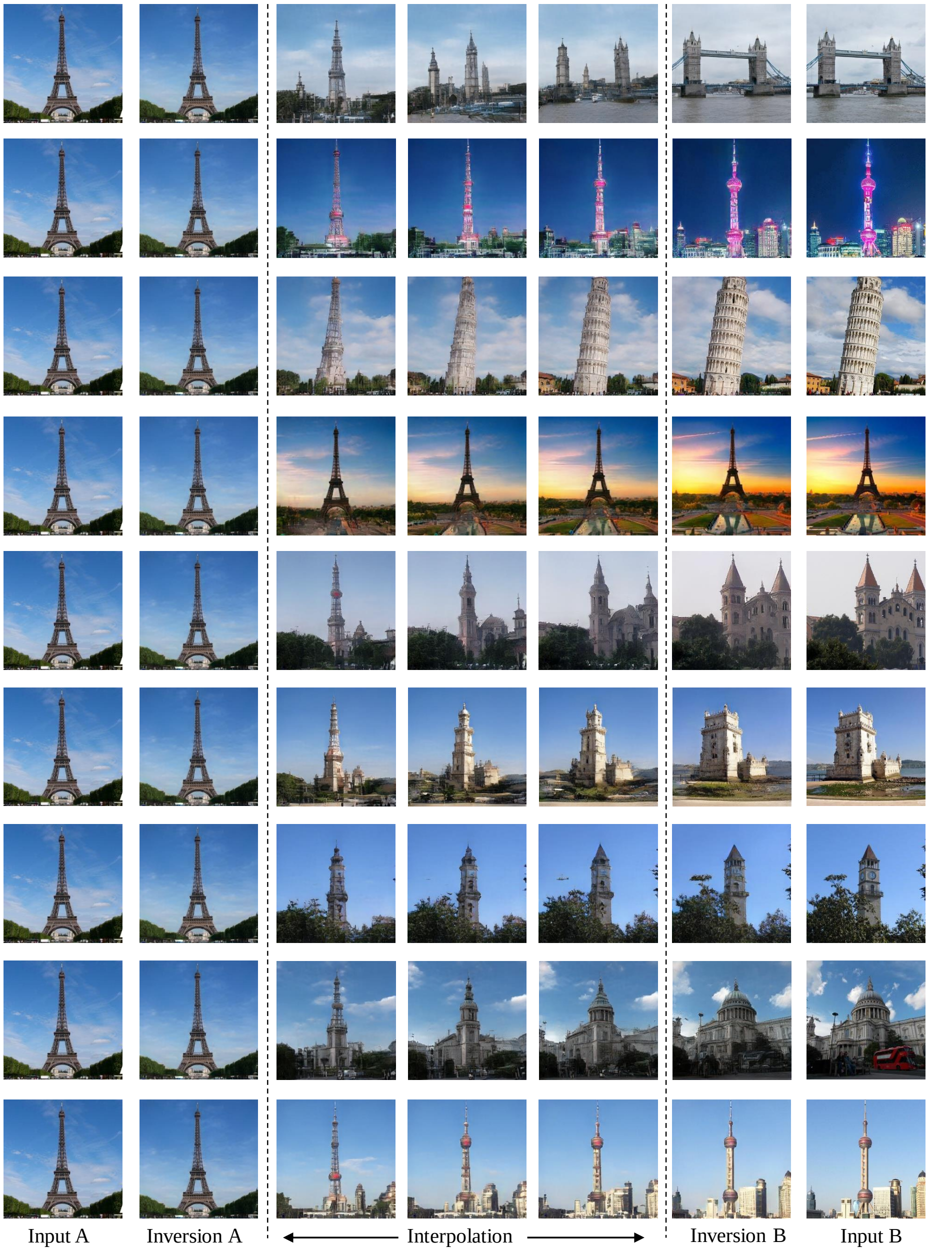


Fig. 6. Tower interpolation results using our *in-domain* GAN inversion method. Zoom in for details.

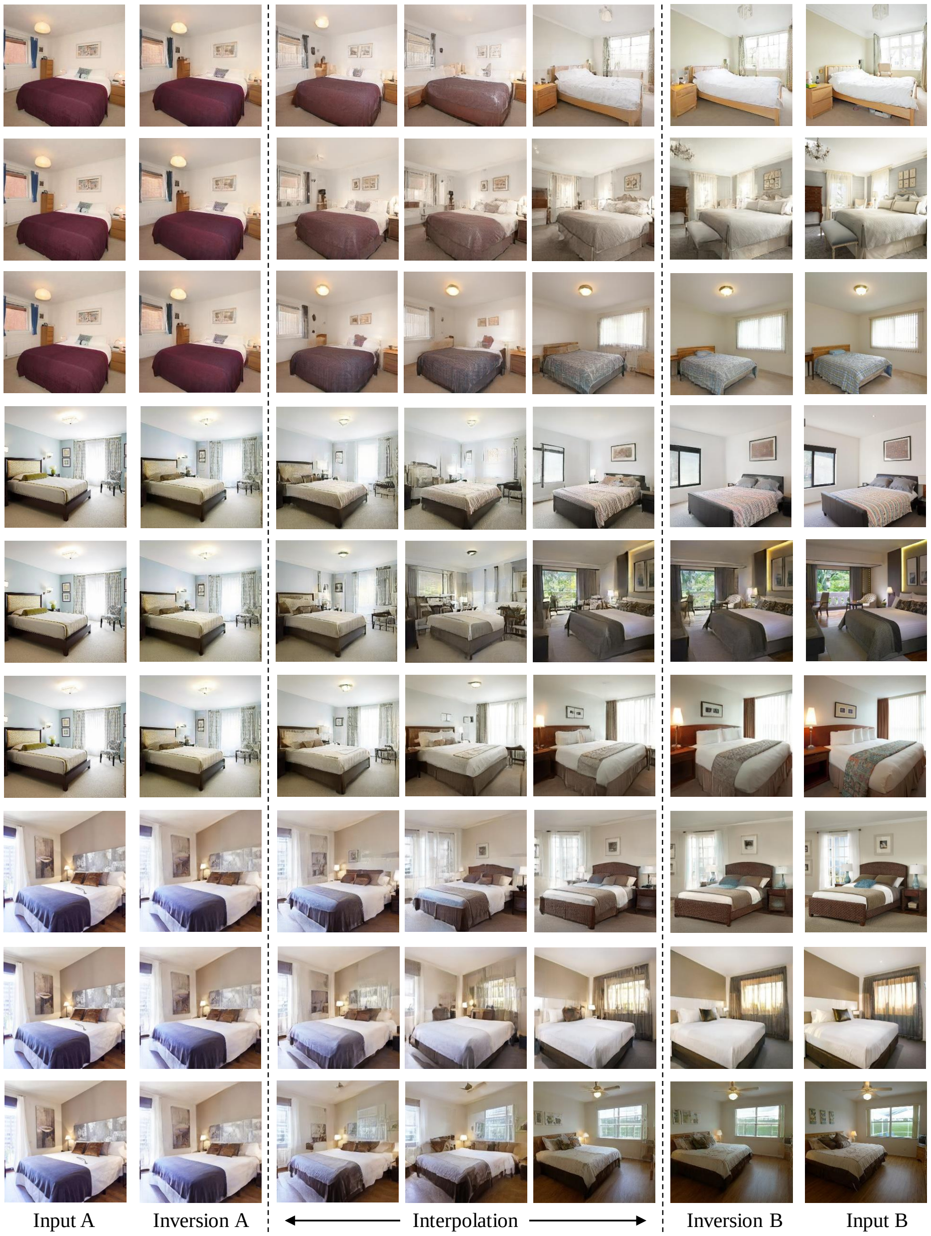


Fig. 7. Bedroom interpolation results using our *in-domain* GAN inversion method. Zoom in for details.

4 Semantic Manipulation

Prior work has shown that a well-trained GAN model is able to encode interpretable semantics inside the latent space [3, 1, 4]. These learned semantics can be further used for real image manipulation together with GAN inversion. In this section, we compare our *in-domain* inversion with Image2StyleGAN [2] on the semantic manipulation task. Results are shown in Fig.8 (face), Fig.9 (tower), and Fig.10 (bedroom). It turns out that we can achieve impressive semantic editing with respect to various attributes, significantly surpassing Image2StyleGAN which usually produces results with artifacts. That is because the code inverted by Image2StyleGAN is not aligned with the rich semantics encoded in the latent space. In other words, only trying to recover the pixel values does not support semantically meaningful image editing. On the contrary, our proposed *in-domain* inversion is able to better reuse the semantic knowledge learned by GANs.

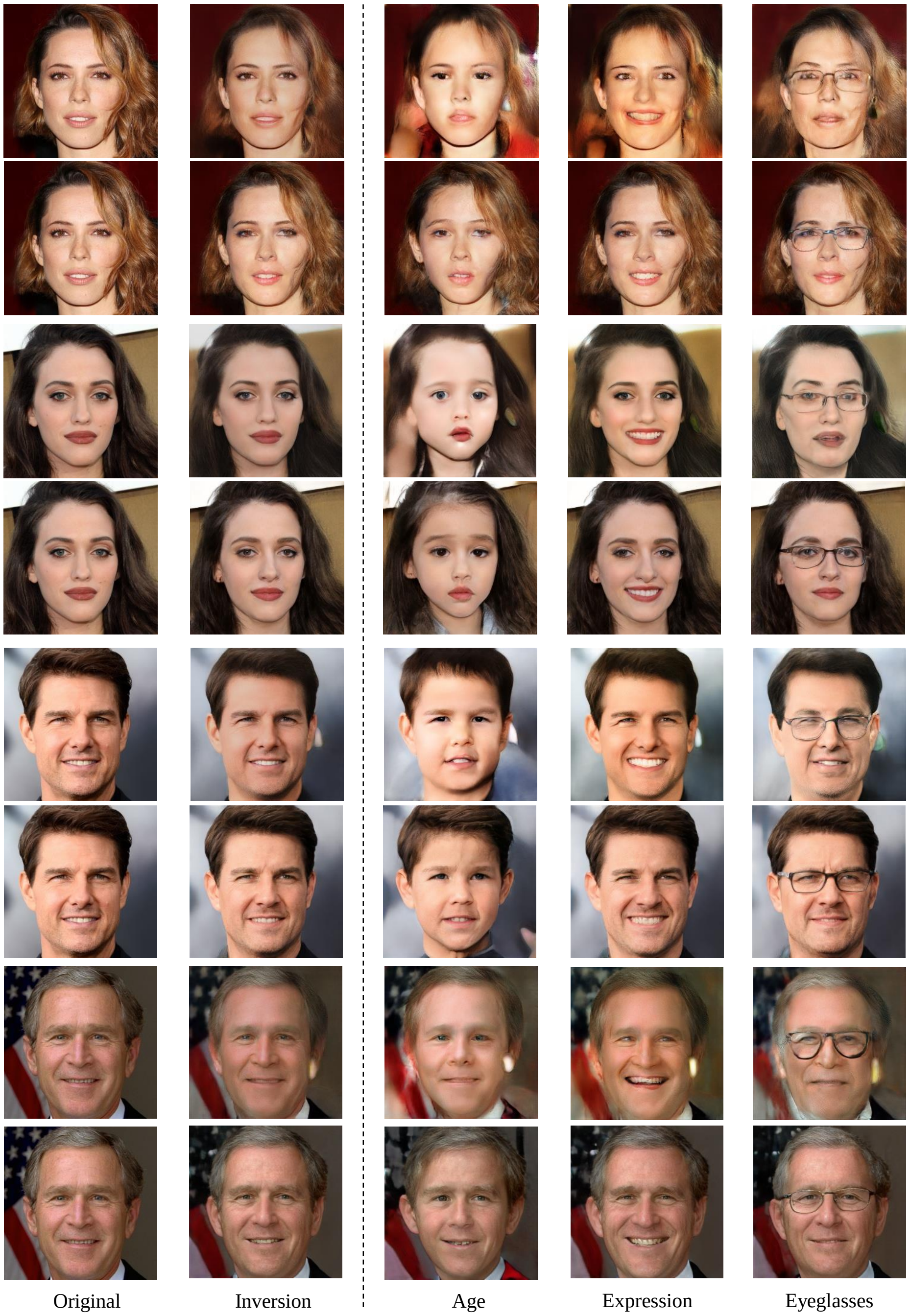


Fig. 8. Comparison results on manipulating face images between Image2StyleGAN [2] and our *in-domain* GAN inversion.

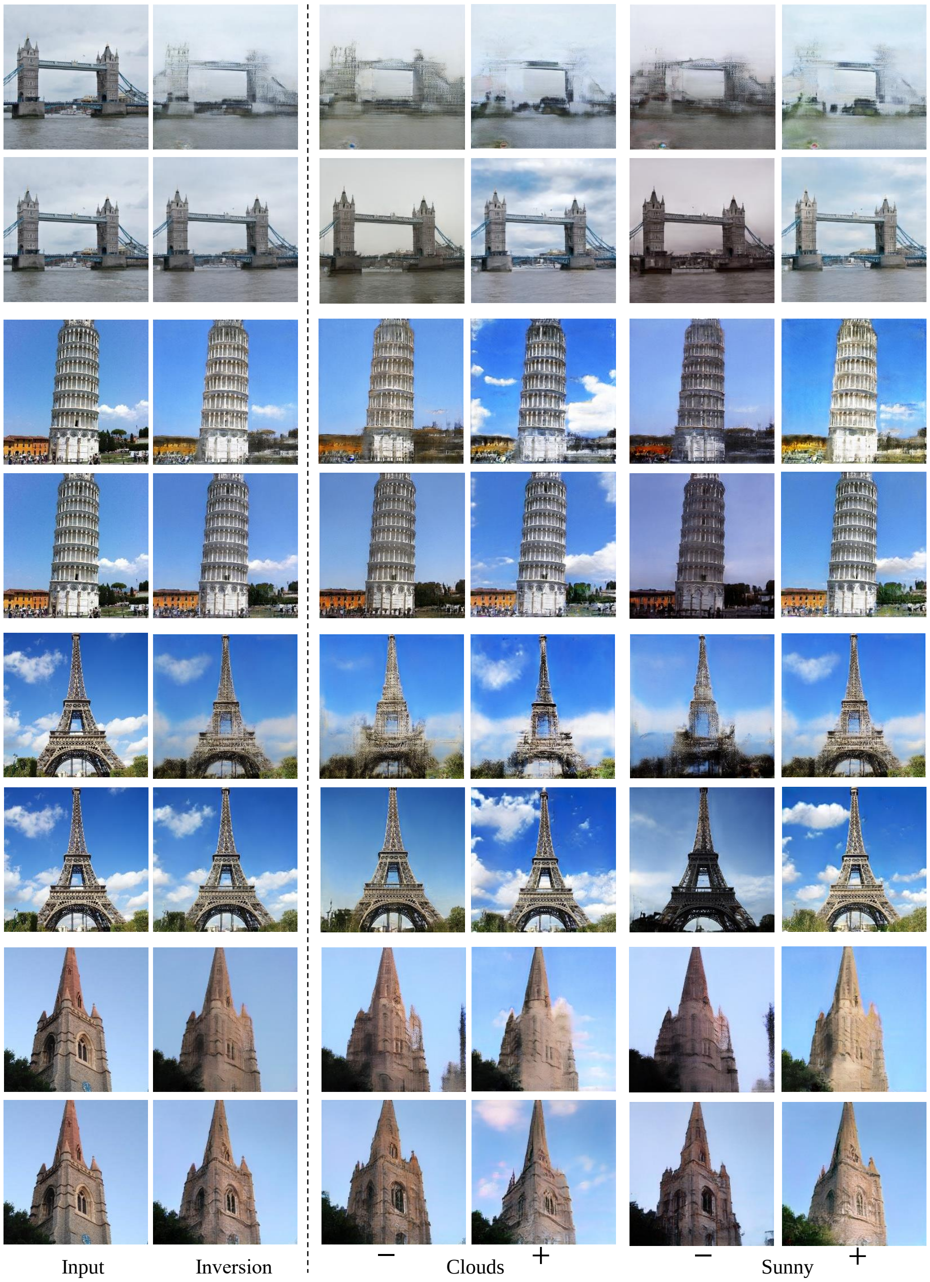


Fig. 9. Comparison results on manipulating tower images between Image2StyleGAN [2] and our *in-domain* GAN inversion.

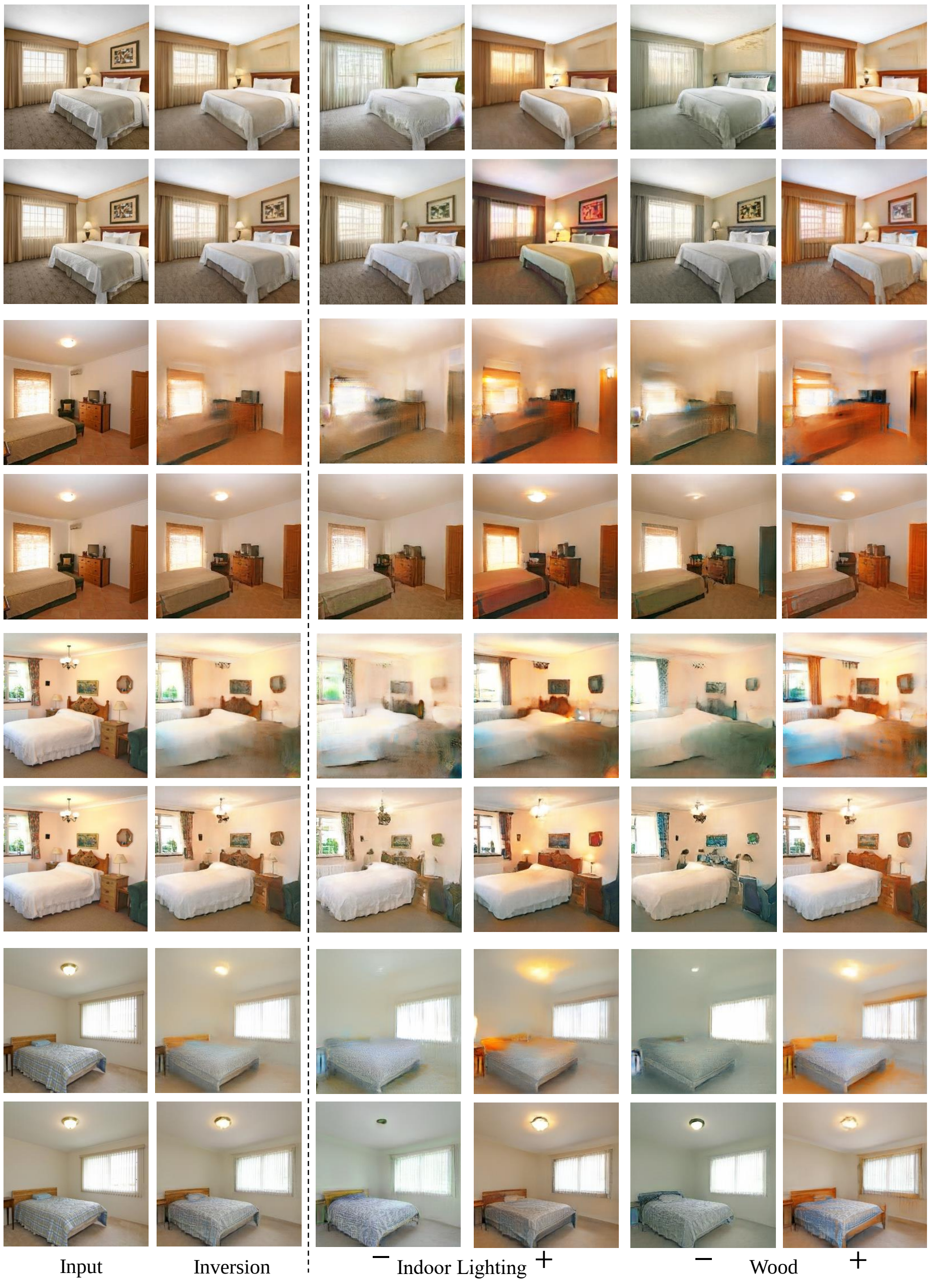


Fig. 10. Comparison results on manipulating bedroom images between Image2StyleGAN [2] and our *in-domain* GAN inversion.

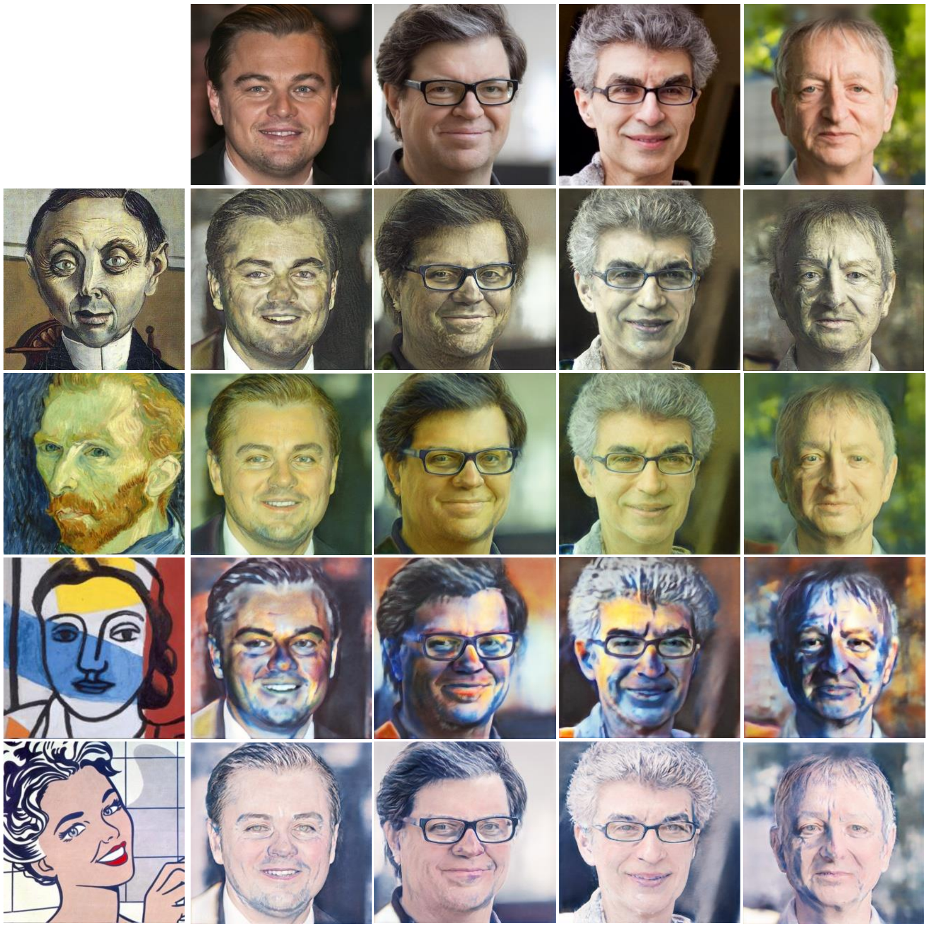


Fig. 11. Style mixing results using our *in-domain* GAN inversion method. First column indicates style images and first row shows content images.

5 Style Mixing

We also evaluate our approach on the style mixing task, which aims at transferring the style of a style image to a content image. For this purpose, we invert both the style image and the content image to layer-wise latent codes. Then, we replace the codes from the last four layers of the content image with those from the style image. Fig.11 shows the mixing results. We can tell that each mixture successfully inherits painting style from the artistic face (first column) yet maintains most details from the real person (first row). This suggests that our *in-domain* inversion manages to convert input images to semantically meaningful latent codes.

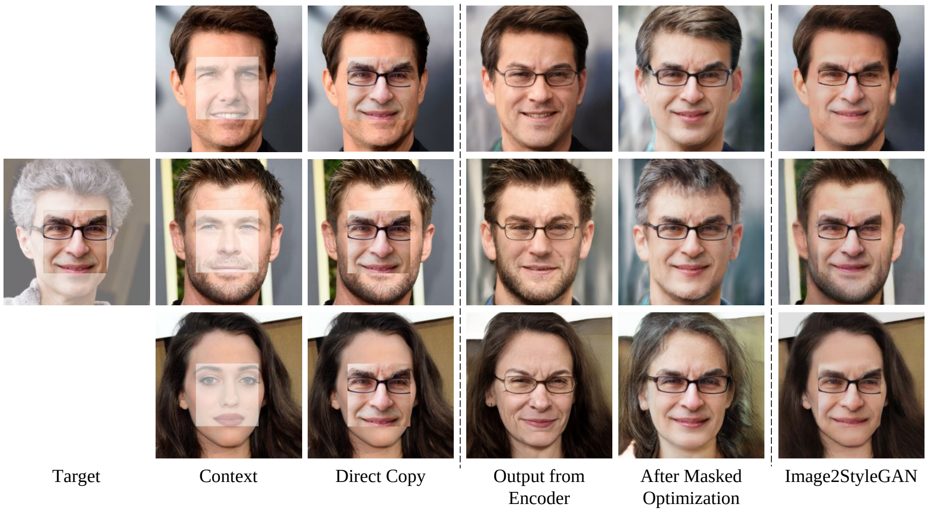


Fig. 12. The reconstruction from the output of our *domain-guided* encoder already has a smooth transition between the target and context. After the masked optimization, the identity of the target is further preserved. As a comparison, Image2StyleGAN [2] fails to produce semantically meaningful image on this task.

6 Semantic Diffusion

In this part, we deeply analyze the semantic diffusion achieved by our *in-domain* inversion.

Implementation Details. Given a target-context image pair, we first crop the wanted part from the target image and then paste it onto the context image. Then, we use our *domain-guided* encoder to infer the latent code for the stitched image. Due to the domain-alignment property of our encoder, the reconstruction from the code can already capture the semantics from both the target patch and its surroundings and further smooth the contents (see the fourth column in Fig.12). With this code as an initialization, we finally perform masked optimization by only using the target foreground region to compute the reconstruction loss. In this way, we are able to not only diffuse the target image to any other context, but also keep the original style of the context image.

Comparison. We show the intermediate results of the semantic diffusion process and compares our approach with the MSE-based inversion approach, Image2StyleGAN [2]. The results are shown in Fig.12, where we have three following observations: (i) The output from our encoder always leads to the reconstruction of a meaningful face and keep most semantics of the inputs (*e.g.*, gender and hair). That is because all codes produced by our encoder are *in-domain*. (ii) The masked optimization is able to preserve the identity information of the target face and further diffuse its style (*e.g.*, skin color) to the surroundings, leading to seamless fusion. This step barely affects the context style (*e.g.* hair style)

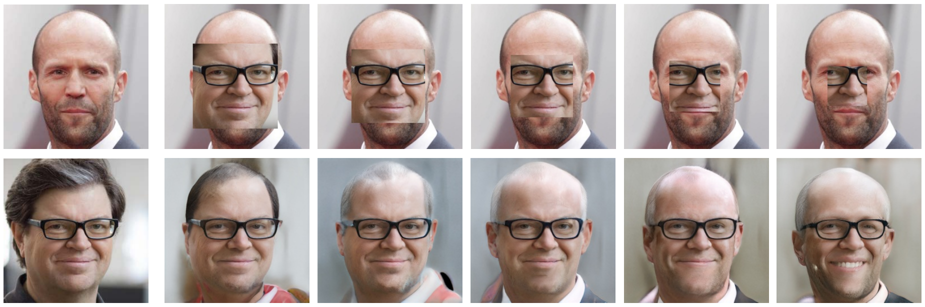


Fig. 13. The effect of crop size on semantic diffusion. Top-left corner shows the context image while bottom-left corner shows the target image. Each remaining column correspond to a different crop size. Top row shows the direct copy-paste results, while bottom row shows the semantic diffusion results.

inherited from the encoder initialization. (iii) Image2StyleGAN fails to produce semantically meaningful faces (*e.g.*, not smooth on the stitch boundary) in the diffusion task since they only focus on the reconstruction of pixel values but not semantics. By contrast, our *in-domain* inversion achieves more satisfying results.

Effect of Crop Size. We further studied the impact of the crop size on semantic diffusion. As shown in Fig.13, we can see that the larger the crop size is (*i.e.* larger reference region from target face), the better the identity information is preserved. For example, on the second column of Fig.13, even hair is transmitted from the target image to the context image since the temples are included in the cropped patch. On the last column, however, the diffused result is no long like the target face at all (*e.g.*, the facial shape and mouth). That is because, during the process of masked optimization, only the foreground patch is used as reference. The surroundings will adaptively change starting from the encoder initialization. Even so, thanks to the *in-domain* property, our approach is still able to complete the entire eyeglasses and generate a smooth diffusion result.

References

1. Jahanian, A., Chai, L., Isola, P.: On the "steerability" of generative adversarial networks. arXiv preprint arXiv:1907.07171 (2019)
2. Rameen, A., Yipeng, Q., Peter, W.: Image2stylegan: How to embed images into the stylegan latent space? In: ICCV (2019)
3. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: CVPR (2020)
4. Yang, C., Shen, Y., Zhou, B.: Semantic hierarchy emerges in deep generative representations for scene synthesis. arXiv preprint arXiv:1911.09267 (2019)
5. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
6. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: ECCV (2016)