A Style-Based GAN Encoder for High Fidelity Reconstruction of Images and Videos: Supplementary Material

Xu Yao[1], Alasdair Newson[1], Yann Gousseau[1], and Pierre Hellier[2]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] InterDigital R&I, France
yaoxu.fdu@gmail.com

1 Ablation study

We conduct an ablation study on the training setup to analyze how each part of the losses contributes to the inversion quality. As shown in Table 1, we compare the inversion quality in the case of removing the the per-pixel loss in eq.(2) (config A), the identity loss in eq.(6) (config B) and the face parsing loss in eq.(7) (config C). In config D, we replace the multi-scale perceptual loss in eq.(3) by a normal LPIPS loss [16], and change its weight λ_1 to scale the loss value to a similar magnitude as before. In config E, we discard the feature prediction branch and generate the inversion with only the latent code. In config F, we use only real images as training data.

As presented in Table 1, we observe a decrease in the perceptual quality in the case of no pixel-wise loss $\mathcal{L}_{mse}$, no identity preservation $\mathcal{L}_{id}$ or no face parsing term $\mathcal{L}_{parse}$. For D, we observe a comparable result on the distortion metrics, but a much higher FID compared to our baseline. E confirms that the feature tensor helps to generate an inversion with better fidelity. F demonstrates that including synthetic data in the training helps improving the perceptual quality of the inversion results. Overall our baseline achieves better perceptual quality and comparable performance on the distortion metrics.

2 Editing

2.1 Choice of K

In this part, we discuss the choice of the feature tensor insertion layer K. As mentioned in the main paper, to perform full editing on the output image, we need to edit the encoded feature tensor $\mathbf{F}$ using eq.(1). One essential condition is that $\mathbf{F}$ remains close to the synthetic feature tensor $\mathbf{G}(\mathbf{w}^{1:K})$, justifying the feature reconstruction loss in eq.(4).

Intuitively, a larger K favors a better reconstruction. However, when K is too large, the feature reconstruction error remains high, which makes it impossible to perform editing using eq.(1). A smaller K favors a lower feature reconstruction error, thus favors better editing results. Meanwhile, K should not be too small,

Table 1. Ablation study on the experimental setup. We conduct experiments on different configurations. For each metric, values indicating lower performance are underlined. Overall our baseline achieves better perceptual quality and comparable performance on the distortion metrics

Configuration	SSIM ↑	PSNR ↑	LPIPS ↓	ID ↑	FID ↓
A. w/o $\mathcal{L}_{mse}$	0.627	23.02	<u>0.242</u>	0.846	19.68
B. w/o $\mathcal{L}_{id}$	0.603	22.65	<u>0.242</u>	<u>0.618</u>	19.26
C. w/o $\mathcal{L}_{parse}$	0.619	22.98	<u>0.248</u>	0.834	<u>20.03</u>
D. w/o $\mathcal{L}_{m\_lpips}$	0.647	24.01	0.056	0.874	<u>22.63</u>
E. w/o feature tensor	<u>0.489</u>	<u>19.44</u>	0.192	<u>0.635</u>	<u>35.28</u>
F. w/o synthetic data	0.644	23.67	0.065	0.873	<u>20.45</u>
G. our baseline	0.641	23.65	0.066	0.867	19.03

otherwise the feature tensor cannot capture enough spatial details. In order to study the corresponding tradeoff, we consider values $K = 4$, $K = 5$, $K = 6$ and $K = 7$. A different model is trained for each configuration.

To analyze each choice, we compare the inversion and style mixing results. Given a source image and a reference image, the style mixing result is generated from the feature tensor of the source image and the latent code of the reference image. Figure 1 shows qualitative results for inversion and style mixing. We observe that using $K = 4$ yields good style mixing effects but low reconstruction quality. Using $K = 6$ or $K = 7$ generates high fidelity reconstruction but the style mixing effects are less obvious. This is because nearly all the information is encoded in the feature tensor, which limits the editing process. Our final choice of $K = 5$ achieves a balance between the inversion quality and the editing capacity.

2.2 Feature Tensor Analysis

We compare the inversion and editing results of using only the latent code and our proposed approach in Figure 2. As can be observed, the inversion result with the feature tensor $\mathbf{G}(\mathbf{F}, \mathbf{w}^{K+1:N})$ captures better the identity and the spatial details compared with the inversion with only the latent code $\mathbf{G}(\mathbf{w})$. We compare also the editing results using our proposed approach $\mathbf{G}(\tilde{\mathbf{F}}, \tilde{\mathbf{w}}^{K+1:N})$ with the editing result generated only from the modified latent code $\mathbf{G}(\tilde{\mathbf{w}})$. Results show that our approach achieves comparable editing capacity while preserving better the spatial details of the original source image.

2.3 Additional Results

We show additional facial attribute editing results in Figure 4. The latent editing directions are computed using InterFaceGAN [11], except the last attribute 'pose', computed with SeFa [12]. SeFa is a method that decomposes the pretrained weights of the generator and further discovers interpretable directions. As can be observed, our method can handle editing of attributes which are controlled by lower layers, such as smile, eyeglasses and pose.

Fig. 1. Choice of the feature tensor insertion layer K. The first column shows the source image (yellow frame) and two reference images for style mixing. In the second to last column, the first row is the inversion results of each configuration, the second and third rows are the style mixing results, generated from the feature tensor of the source image and the latent code of the reference image. Using $K = 4$ yields good style mixing effects but low reconstruction quality. Using $K = 6$ or 7 encodes nearly all the information in the feature tensor, which is limiting for editing. Our choice of $K = 5$ holds a balance between the editing capacity and the reconstruction quality.

We also present additional editing result for StyleGAN2 pre-trained on the car domain in Figure 5 and StyleGAN2-Ada pre-trained on cat domain in Figure 6 and on dog domain in Figure 7. As the attributes of the car dataset [8] and the cat/dog dataset [3] are not annotated, all the latent editing directions are computed with SeFa [12]. Please note that the directions discovered by SeFa are not necessarily disentangled. Overall, our model generates better inversion results and yields satisfying editing results.

3 Inversion

We show additional visual results for the inversion of StyleGAN2 pre-trained on face domain in figs. 8 to 12. We compare our model against the optimization based method [1], state-of-the-art encoder based methods [10,13,2,14] and a hybrid method [17]. As can be observed, reconstructions using our framework are visually more faithful and zoom-in patches show that they exhibit much more details and sharpness.

We show more inversion results for StyleGAN3 [6] pre-trained on FFHQ [7]. Compared with StyleGAN2, the architecture of StyleGAN3 has several important changes. First, the input tensor passed into the generator is no longer con-

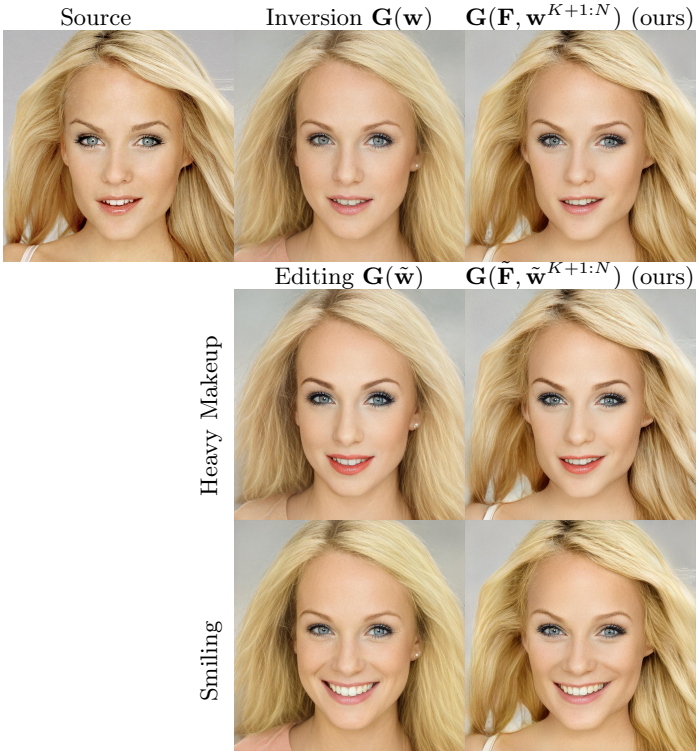

Fig. 2. Feature tensor analysis. The image on the top left is the source image. The second image of the first row is the inversion result with only the latent code $\mathbf{G}(\mathbf{w})$, the third one is the inversion result with the feature tensor $\mathbf{G}(\mathbf{F}, \mathbf{w}^{K+1:N})$. The second and third rows show the editing results of attribute 'Heavy Makeup' and 'Smiling'. $\mathbf{G}(\tilde{\mathbf{w}})$ is the editing result corresponding to the inversion $\mathbf{G}(\mathbf{w})$, and $\mathbf{G}(\tilde{\mathbf{F}}, \tilde{\mathbf{w}}^{K+1:N})$ corresponds to our proposed editing approach – edit both the latent code and the feature tensor. As can be observed, our approach achieves comparable editing capacity while preserving better the spatial details of the original source image.

stant, but synthesized from the latent code. The spatial size of the input tensor is increased from 4×4 to 36×36. Additionally, the noise inputs are discarded. As shown in Figure 13, despite the architectural changes, our proposed encoder still yields satisfying inversion results.

4 Video results

We provide qualitative video results on inversion and editing in *this link*. In each subfolder, 'inversion.avi' is the inversion result, 'edit_attribute.avi' is the latent editing result. *Please open the videos to better visualize the results.* The videos for evaluation are collected from the FILMPAC library [5]. The inversion and editing results are generated using the video manipulation pipeline proposed in [15]. In

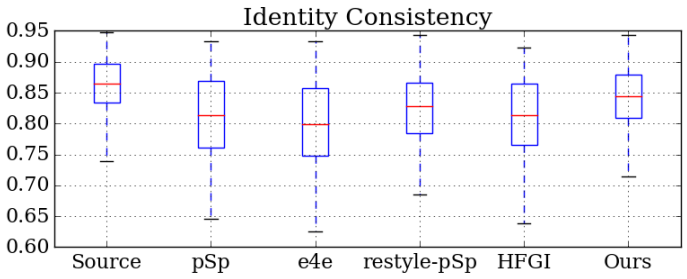

Fig. 3. Identity consistency of video inversion. For each method, we compute the proposed metric *identity consistency* for each inverted video and plot the results in a box-plot. Our averaged identity consistency is the closest to that of the source videos.

each video, the first row shows the original video, the result of pSp [10] and that of e4e [13]. The second row shows the result of restyle [2], that of HFGI [14] and our result. Overall our model generates a more consistent inversion, that further favors consistent editing.

Inversion Consistency Additionally, to evaluate the consistency of the inversion, we propose a new metric *Identity Consistency*, which refers to the averaged identity similarity between the reconstructed frame $\tilde{\mathbf{x}}_i$ and frame $\tilde{\mathbf{x}}_0$ along a video sequence:

$$IC = \frac{1}{N}\sum_{i=1}^{N}\langle \mathbf{R}(\tilde{\mathbf{x}}_i), \mathbf{R}(\tilde{\mathbf{x}}_0)\rangle, \tag{1}$$

where $\mathbf{R}$ is the pre-trained ArcFace[4] network. We compute this metric for our encoder and state-of-the-art methods for video inversion on RAVDESS [9]. From this dataset we sample randomly 120 videos as evaluation data. For each method, we perform the inversion and compute this metric on each inverted video and present the results with a box-plot. Figure 3 shows that the averaged identity consistency of our inversion is the closest to that of the source, which proves the stability of our inversion.

Fig. 4. Editing on face domain. We show additional facial attribute editing results. The latent editing directions are computed using InterFaceGAN [11], except the last attribute 'pose', computed with SeFa [12].

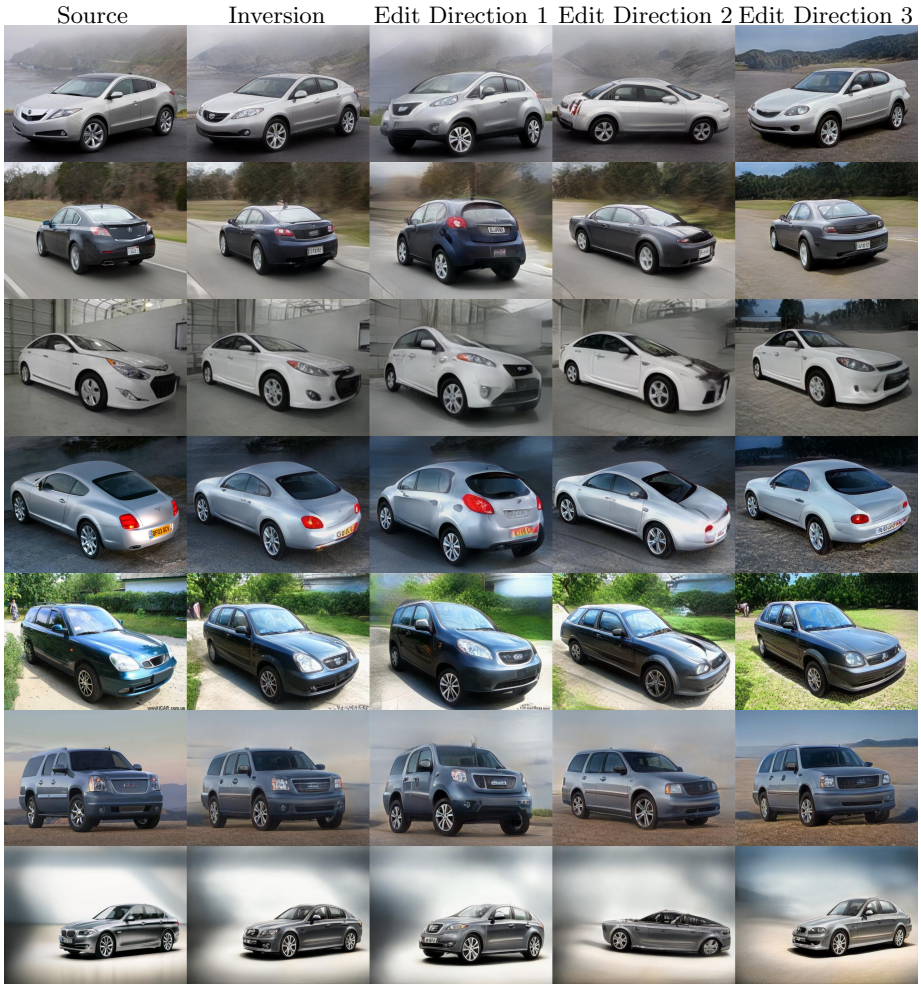

Fig. 5. Editing on car domain. We show latent space editing results on car domain. We compute the latent editing directions with SeFa [12]. The first column is the source image, second column is our inversion result, the third to last column correspond to the semantic directions found with SeFa [12]. Our model yields satisfying editing results on car domain.

Fig. 6. Editing on cat domain. We show latent space editing results on cat domain. We compute the latent editing directions with SeFa [12]. The first column is the source image, second column is our inversion result, the third to last column correspond to the semantic directions found with SeFa [12]. Our model yields satisfying editing results on cat domain.

Fig. 7. Editing on dog domain. We show latent space editing results on dog domain. We compute the latent editing directions with SeFa [12]. The first column is the source image, second column is our inversion result, the third to last column correspond to the semantic directions found with SeFa [12]. Our model yields satisfying editing results on dog domain.

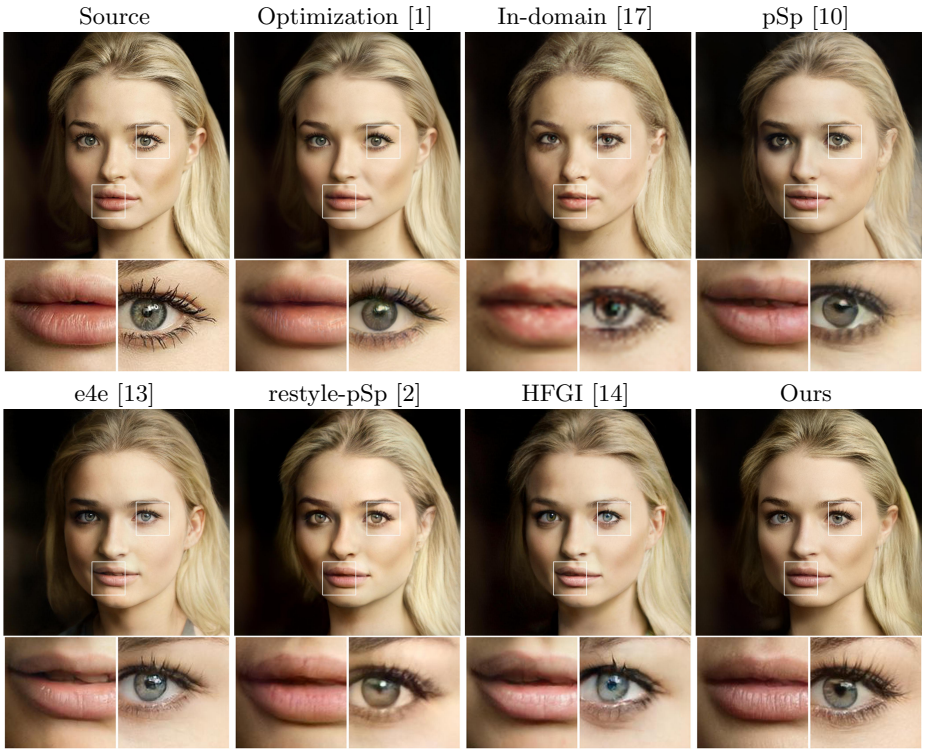

Fig. 8. Inversion on face domain. We compare our model against state-of-the-art GAN inversion methods [1,17,10,13,2,14] for the inversion of StyleGAN2 pre-trained on face domain. Reconstructions using our framework are visually more faithful and zoom-in patches show that they exhibit much more details and sharpness.

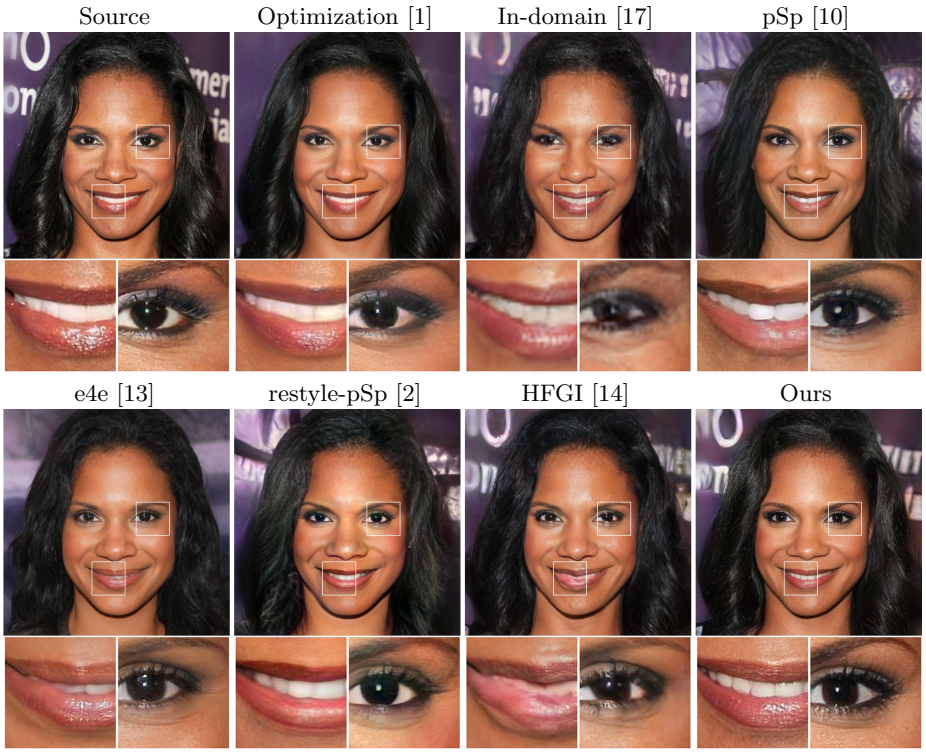

Fig. 9. Inversion on face domain. We compare our model against state-of-the-art GAN inversion methods [1,17,10,13,2,14] for the inversion of StyleGAN2 pre-trained on face domain. Reconstructions using our framework are visually more faithful and zoom-in patches show that they exhibit much more details and sharpness.

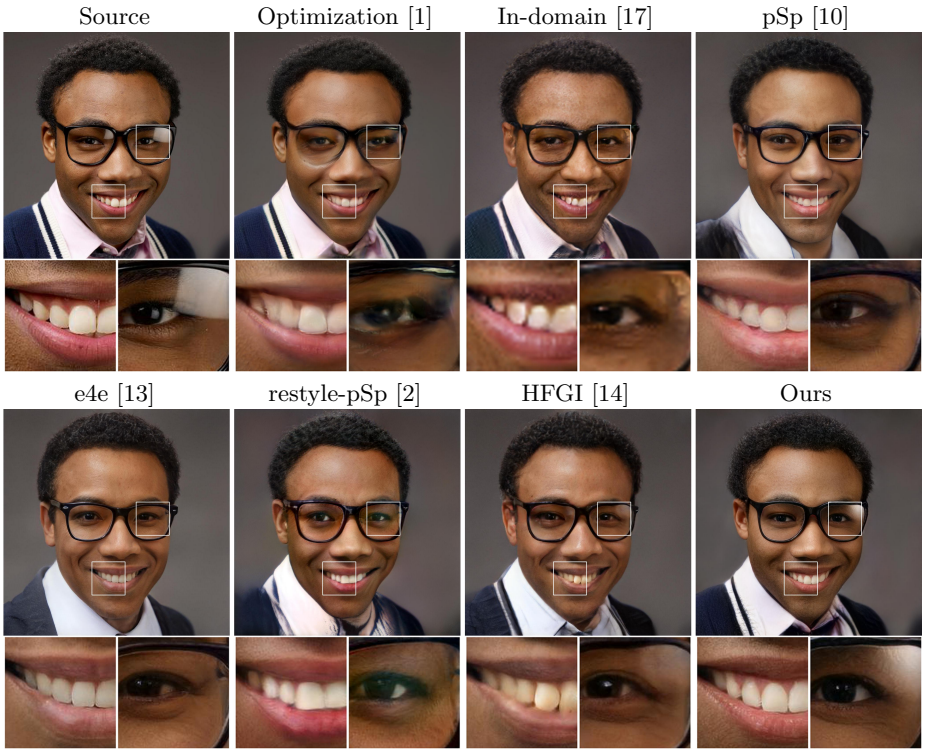

Fig. 10. Inversion on face domain. We compare our model against state-of-the-art GAN inversion methods [1,17,10,13,2,14] for the inversion of StyleGAN2 pre-trained on face domain. Reconstructions using our framework are visually more faithful and zoom-in patches show that they exhibit much more details and sharpness.

Fig. 11. Inversion on face domain. We compare our model against state-of-the-art GAN inversion methods [1,17,10,13,2,14] for the inversion of StyleGAN2 pre-trained on face domain. Reconstructions using our framework are visually more faithful and zoom-in patches show that they exhibit much more details and sharpness.

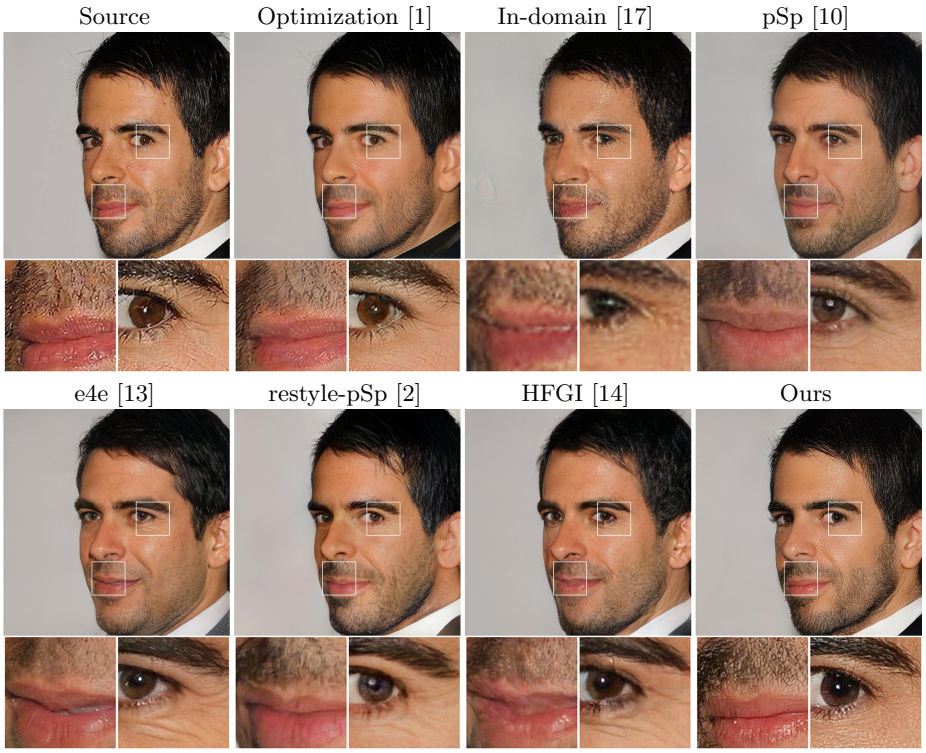

Fig. 12. Inversion on face domain. We compare our model against state-of-the-art GAN inversion methods [1,17,10,13,2,14] for the inversion of StyleGAN2 pre-trained on face domain. Reconstructions using our framework are visually more faithful and zoom-in patches show that they exhibit much more details and sharpness.

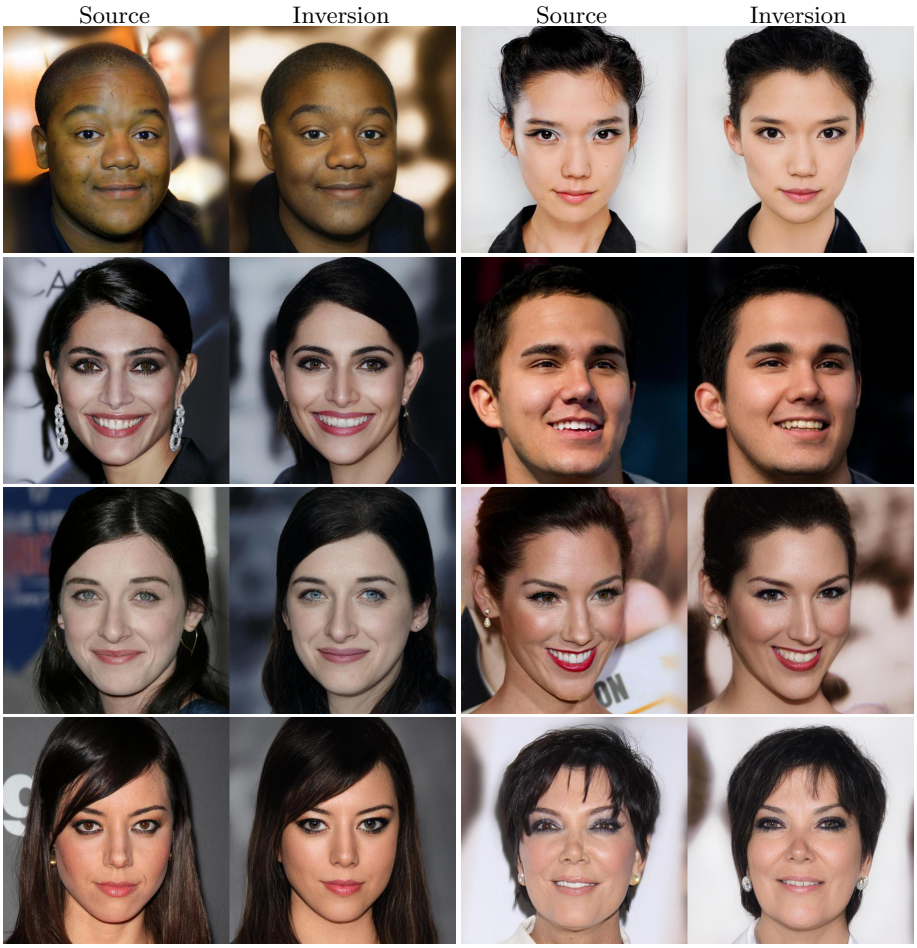

Fig. 13. Inversion of StyleGAN3 pretrained on face domain. We show preliminary inversion results of the 3rd generation of StyleGAN [6] on face domain. Compared with StyleGAN2, the architecture of StyleGAN3 has several important changes. Despite the architectural changes, our method still yields satisfying inversion results

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8296–8305 (2020)
2. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6711–6720 (2021)
3. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8188–8197 (2020)
4. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
5. FILMPAC: Filmpac footage boutique library. `https://filmpac.com` (2017)
6. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: Proc. NeurIPS (2021)
7. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
8. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
9. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one **13**(5), e0196391 (2018)
10. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2287–2296 (2021)
11. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9243–9252 (2020)
12. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1532–1540 (2021)
13. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) **40**(4), 1–14 (2021)
14. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. arXiv preprint arXiv:2109.06590 (2021)
15. Yao, X., Newson, A., Gousseau, Y., Hellier, P.: A latent transformer for disentangled face editing in images and videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13789–13798 (2021)
16. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
17. Zhu, J., Shen, Y., Zhao, D., Zhou, B.: In-domain gan inversion for real image editing. In: European Conference on Computer Vision. pp. 592–608. Springer (2020)