High-fidelity GAN Inversion with Padding Space
– *Supplementary Material* –

Qingyan Bai[1]*, Yinghao Xu[2]*, Jiapeng Zhu[3], Weihao Xia[4],
Yujiu Yang[1]†, and Yujun Shen[5]

[1]Shenzhen International Graduate School, Tsinghua University [2] CUHK
[3] HKUST [4] University College London [5] ByteDance Inc.
bqy20@mails.tsinghua.edu.cn, xy119@ie.cuhk.edu.hk
{jengzhu0, xiawh3}@gmail.com
yang.yujiu@sz.tsinghua.edu.cn, shenyujun0302@gmail.com

A Overview

This supplementary material is organized as follows. Sec. B describes the detailed architecture of our encoder. Sec. C verifies how the regularization loss, $\mathcal{L}_{reg}$, helps improve the editing performance. Sec. D further validates the property of the proposed padding space with interpolation results, as the supplement to Sec. 4.3 of the submission. Sec. E provides more visual results on inversion, face blending, and customized manipulation.

B Encoder Structure

Tab. S1 provides the detailed architecture of our encoder, by taking a 18-layer StyleGAN [5,6] generator as an instance. Note that input images are resized to 256×256 at first. Features are first extracted from the backbone and refined by FPN. Then style latent codes of $\mathcal{W}+$ space are obtained with Map2Style borrowed from [9]. Paddings are obtained by convolving the 512×32^2 FPN feature map with the Resblocks and 1×1 convolutions (denoted as "Padding Extracting Convolutions"). The "Style Codes" term indicates which layer of StyleGAN the style latent codes will be sent to. Additionally, we equip the Resblocks with squeeze and excitation block [10] to enhance them.

C Ablation Study on Regularization Loss

Recall that we introduce the regularization loss $\mathcal{L}_{reg}$ in Eq.(5) of the submission to enhance the encoder in editing ability. In order to achieve better editing performance, this regularization is applied to encourage the inverted code in $\mathcal{W}+$ space to be subject to the native latent distribution.

* Equal contribution.
† Corresponding author.

Table S1. Encoder structure based on ResNet-IR-50 [2,3]. The numbers in brackets indicate the dimension of features at each level. Paddings are obtained by convolving the 512×32^2 FPN feature map with the Resblock and 1×1 convolutions (denoted as "Padding Extracting Convolutions")

Stage	Backbone	Output Shape	FPN	Map2Style	Style Codes	Padding Extracting Convolutions
input	–	3×256^2				
conv_1	3×3, 64 stride 1, 1	64×256^2				
res_2	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 3$	64×128^2				
res_3	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 4$	128×64^2	512×64^2	11×512	Layer 8-18	
res_4	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 14$	256×32^2	512×32^2	4×512	Layer 4-7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$ & $\begin{bmatrix} 1\times1, 512 \end{bmatrix} \times 4$
res_5	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 3$	512×16^2	512×16^2	3×512	Layer 1-3	

In this section we ablate $\mathcal{L}_{reg}$ on editing tasks. We separately learn two proposed encoders with or without $\mathcal{L}_{reg}$ and use them to invert the input faces. Then off-the-shelf semantic directions from [11, 14] are adopted to edit the inversion results. As demonstrated in Fig. S1, during editing the model trained without $\mathcal{L}_{reg}$ produces inferior results such as failed editing for eye-glasses, blurred teeth for smile, and unnatural expression for bigger eyes. While models trained with L_{reg} have better performance in editing but show slightly inferior performance in reconstruction.

D Property of Padding Space

Recall that we conduct interpolation experiments in Sec 4.3 of the submission. Here we additionally provide the interpolation results on the FFHQ test set (the spared last 5k images) and LSUN Bedroom test set to validate the property of the padding space ($\mathcal{P}$ space).

Firstly, we fixed the style latent codes as the average and interpolate the paddings between the fixed constants and coefficients extracted by the encoder. From Fig. S2 and Fig. S3 we can conclude that during interpolation the style information such as face details and color style is maintained. While the pose and contour keep changing from the average to the input image. Then as in Sec 4.3 of the submission, we perform interpolation between two inversion results. Specifically, we first invert two real images A and B to $\mathcal{P}$ and $\mathcal{W}+$ space. Then we fix one of the paddings or latent codes and interpolate the other. As in Fig. S4, when padding coefficients are kept to be fixed, the style varies smoothly when the latent code changes. Also when style latent codes are fixed, the spatial information changes from the image A to B while the image style stays fixed. In summary, these results indicate that the paddings encode spatial information such as the contour of the face, background, and pose.

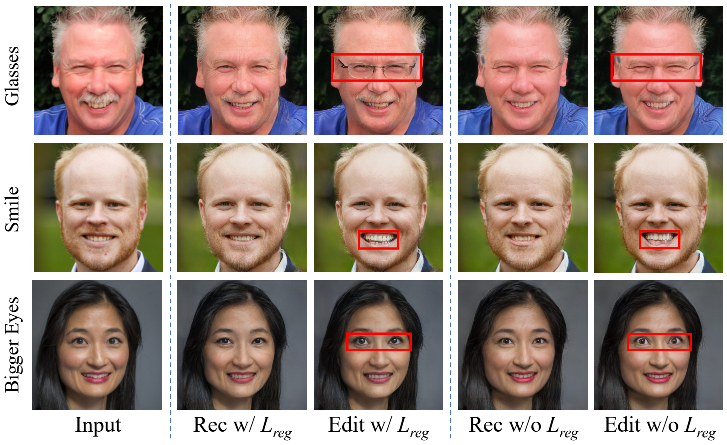

Fig. S1. Ablation study on L_{reg}. Models trained with L_{reg} have better performance in editing while show slightly inferior performance in reconstruction

E Additional Results

In this section, we provide more visualization results of inversion, face blending and customized editing.

E.1 Inversion

Here we qualitatively compare our inversion results with the ones of ALAE [8], IDInvert [15], pSp [9], e4e [12], and Restyle [1]. Fig. S5 and Fig. S6 respectively demonstrate inversion results on FFHQ [5] and LSUN [13] test set, from which we conclude that our method performs better in reconstruction of spatial details. We additionally validate the effectiveness of the proposed padding space on StyleGAN1 [5]. Fig. S7 demonstrate the inversion results of our method and the baseline pSp with pretrained StyleGAN1 on CelebA-HQ [4, 7] test set.

E.2 Face Blending

Recall that our method enables face blending by simply swapping the style latent codes of two input faces. More face blending results are provided in in Fig. S8. We can conclude that the paddings encode spatial information such as the face pose and contour while style latent codes encode face details.

E.3 Customized Editing

Recall that we define semantic directions with paired images and edit the inversion results with them in Sec 4.4 of the submission. Here we provide more customized editing results in Fig. S9. We can conclude that the semantic direction defined

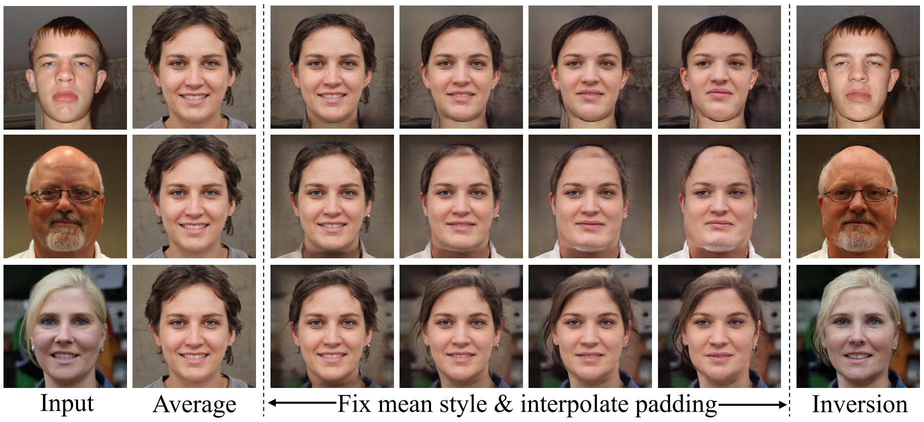

Fig. S2. Analysis of the padding effect on FFHQ [5] test set. We fix the latent code as the statistical average and interpolate the padding from the fixed constants in the generator to the coefficients specifically learned for inversion. It verifies that padding encodes the spatial information

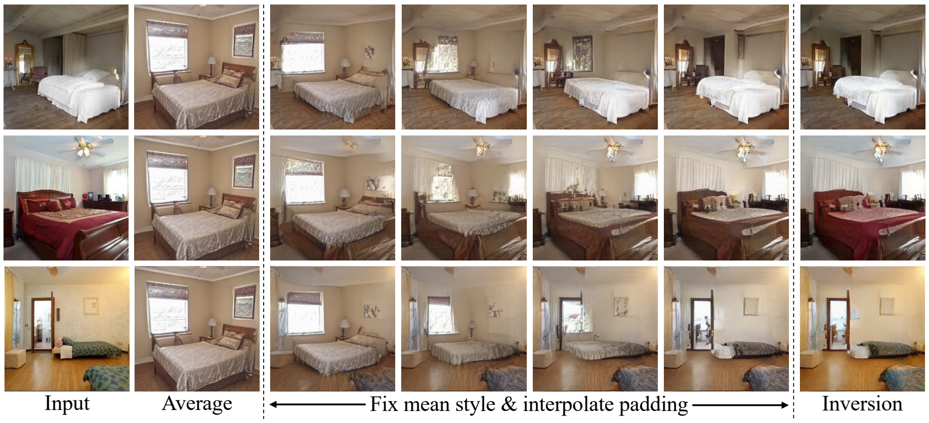

Fig. S3. Analysis of the padding effect on LSUN Bedroom [13]. We fix the latent code as the statistical average and interpolate the padding from the fixed constants in the generator to the coefficients specifically learned for inversion

by one pair of images (A and A') can be applied to arbitrary samples (such as B and C) to precisely edit them.

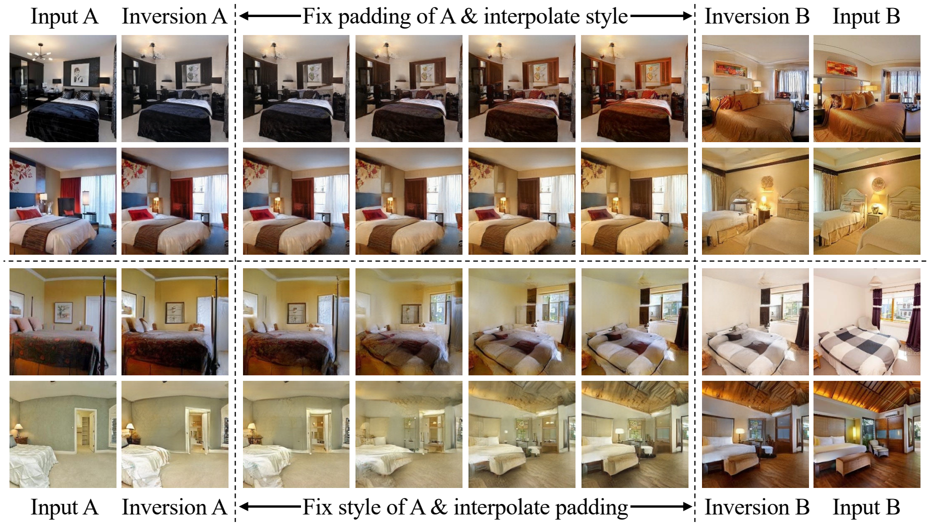
Input A
Inversion A
Fix padding of A & interpolate style
Inversion B
Input B
Input A
Inversion A
Fix style of A & interpolate padding
Inversion B
Input B

Fig. S4. Analysis of the extended inversion space on LSUN Bedroom [13]. We perform interpolation both in the latent space and in the padding space

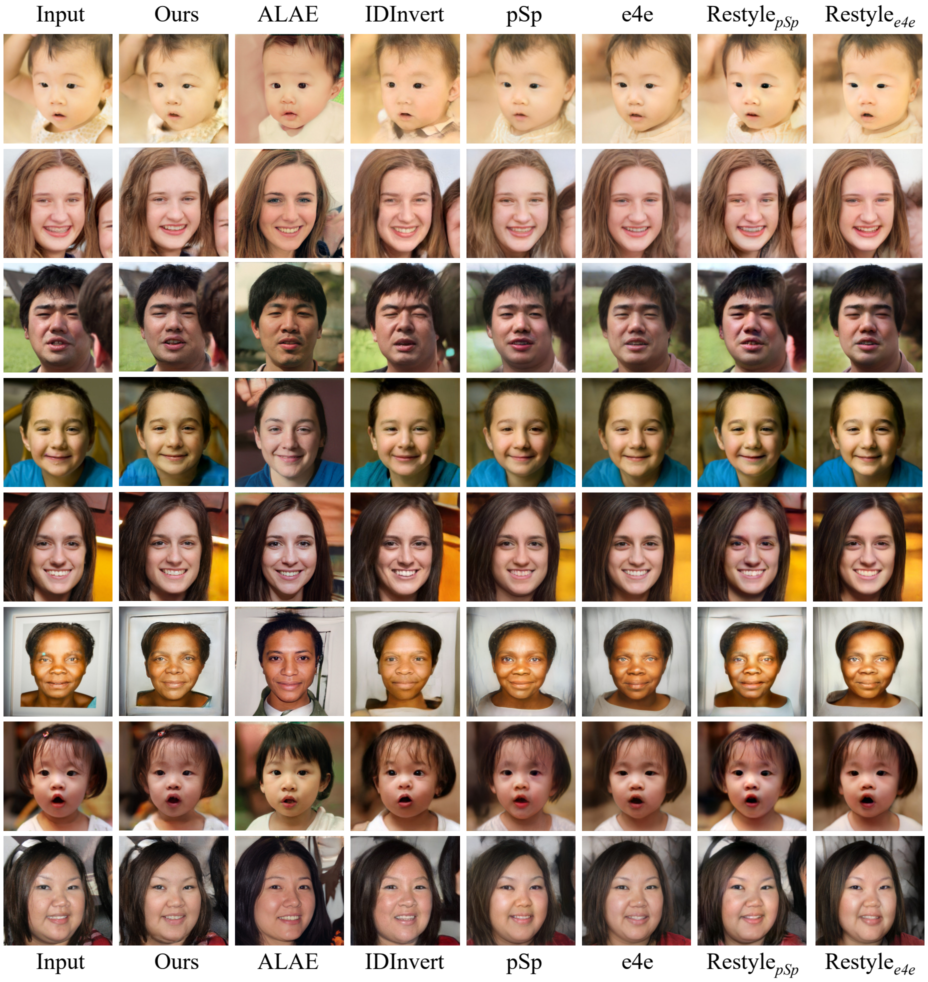

Fig. S5. Additional inversion results of various methods on FFHQ [5] test set. Note that our method perform better when reconstructing out-of-distribution spatial details such as the arm, secondary face, background, hairpins, and head-wears

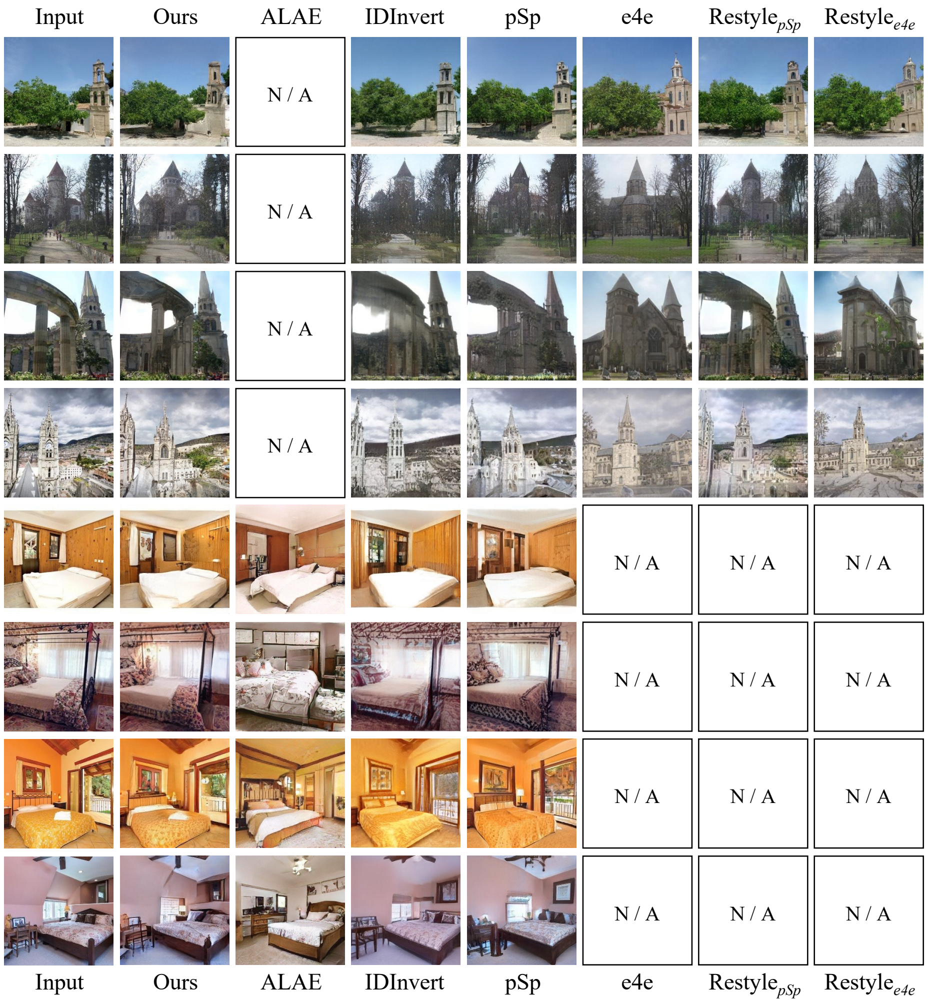

Fig. S6. Additional inversion results of various methods on LSUN [13] Church and Bedroom test set. Our method perform better when reconstructing spatial details such as and holes and pillows. N/A indicates the pretrained model is not available

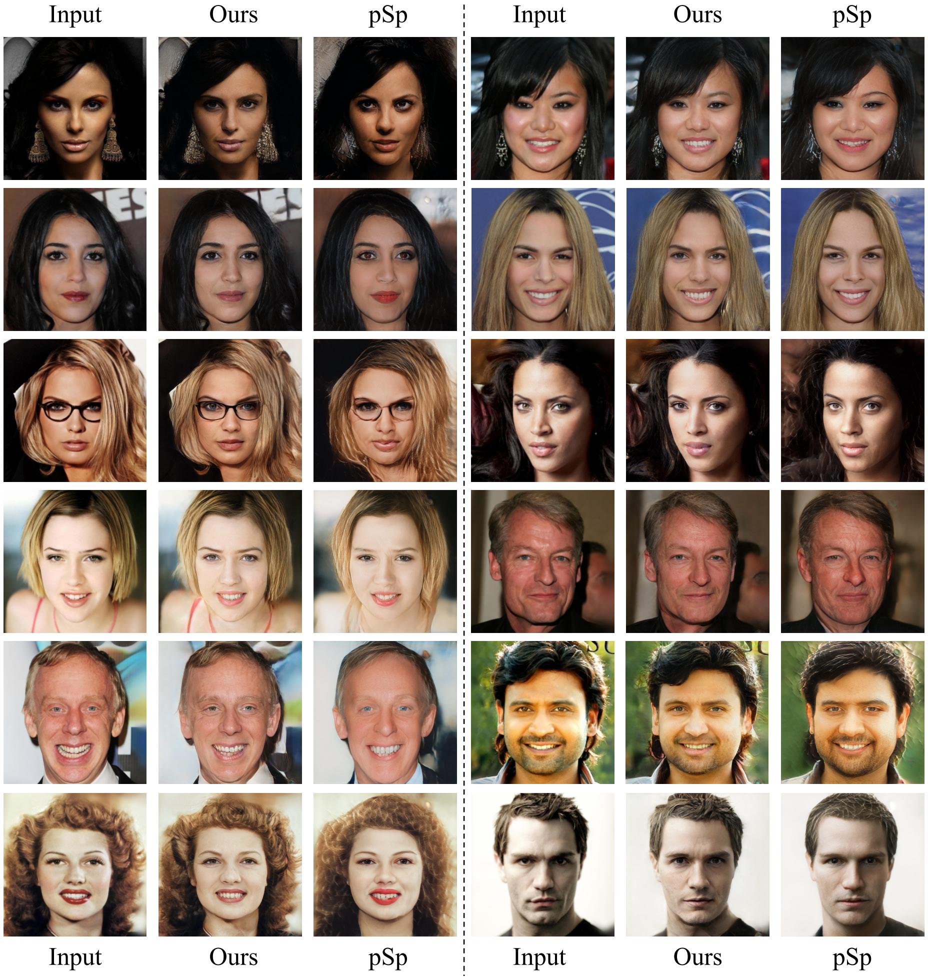

Fig. S7. Additional inversion comparisons between our method and the baseline pSp [9] on CelebA-HQ [4, 7] test set with pretrained StyleGANv1. Our method has better performance in reconstruction of spatial details such as earring, background, and haircut

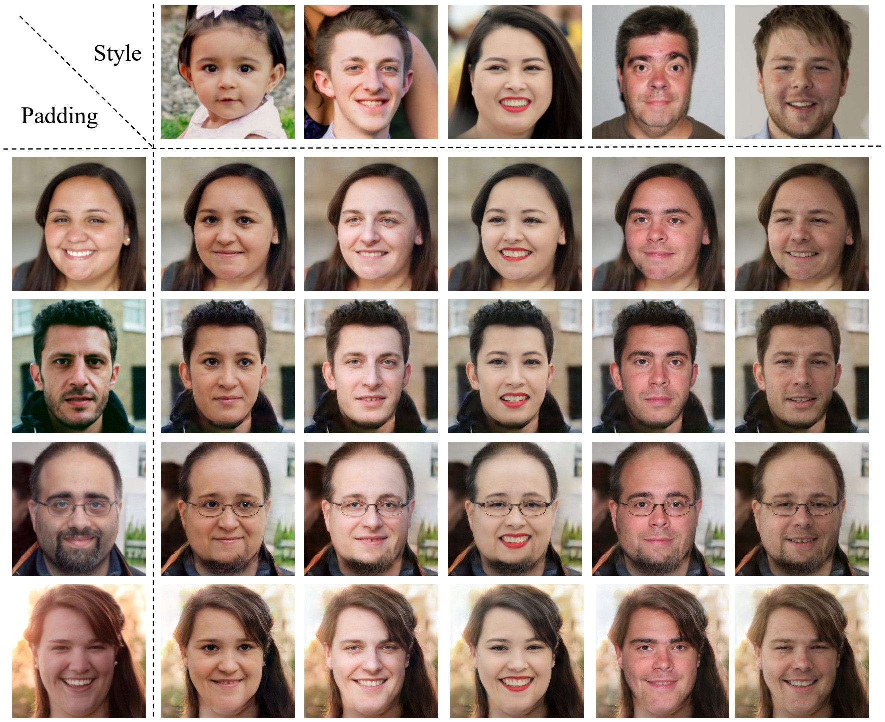

Fig. S8. Additional face blending results by combining different paddings and style latent codes, as described in Sec 4.3 of the submission. We can conclude that the paddings encode spatial information such as face pose and contour while style latent codes encode face details

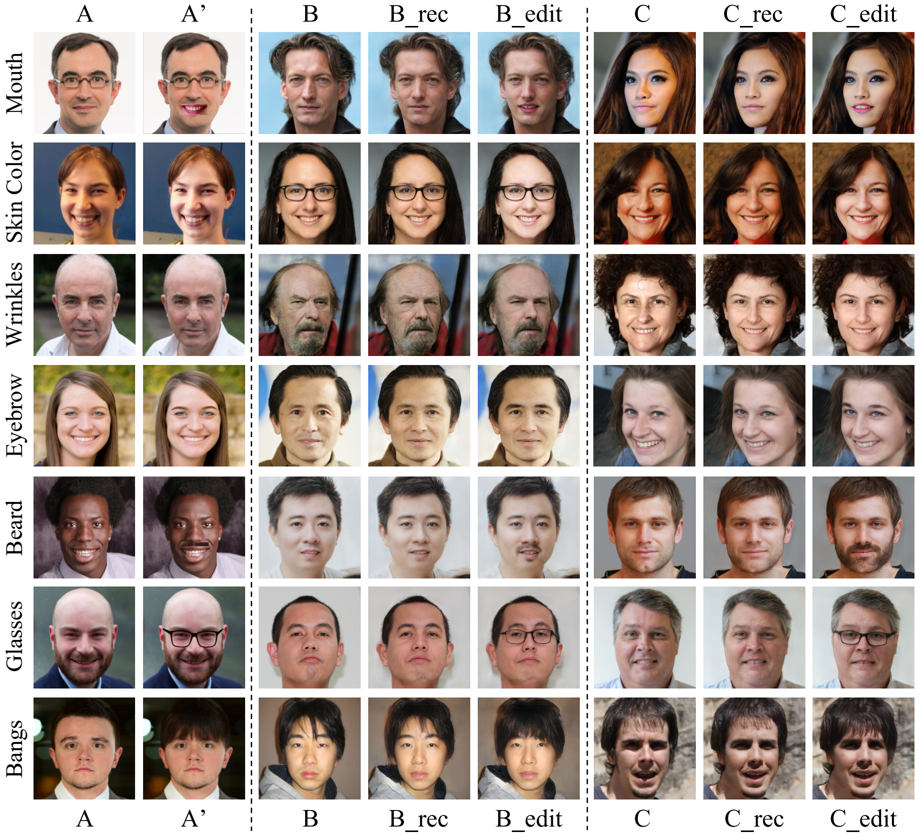

Fig. S9. Additional results of customized editing with one pair of images. The semantic direction defined by one pair of images (A and A') can be applied to arbitrary samples (like B and C) to precisely edit them. "rec" denotes the inversion result of the input. Among the attributes, glasses and bangs are spatial-related thus we edit them only in the padding space and others are style-related and are edited in $\mathcal{W}+$ space, as described in Sec 4.4 of the submission

References

1. Alaluf, Y., Patashnik, O., Cohen-Or, D.: ReStyle: A residual-based StyleGAN encoder via iterative refinement. In: Int. Conf. Comput. Vis. (2021)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: Int. Conf. Learn. Represent. (2018)
5. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019)
6. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
7. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Int. Conf. Comput. Vis. (2015)
8. Pidhorskyi, S., Adjeroh, D.A., Doretto, G.: Adversarial latent autoencoders. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
9. 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: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
10. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 421–429. Springer (2018)
11. Shen, Y., Yang, C., Tang, X., Zhou, B.: InterFaceGAN: Interpreting the disentangled face representation learned by GANs. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
12. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for StyleGAN image manipulation. ACM Trans. Graph. (2021)
13. 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)
14. Zhu, J., Feng, R., Shen, Y., Zhao, D., Zha, Z., Zhou, J., Chen, Q.: Low-rank subspaces in GANs. In: Adv. Neural Inform. Process. Syst. (2021)
15. Zhu, J., Shen, Y., Zhao, D., Zhou, B.: In-domain GAN inversion for real image editing. In: Eur. Conf. Comput. Vis. (2020)