IntereStyle: Encoding an Interest Region for Robust StyleGAN Inversion

Seung-Jun Moon[1] and Gyeong-Moon Park[1,2]

[1] KLleon Tech. seungjun.moon@klleon.io
[2] Kyung Hee University
gmpark@khu.ac.kr

A Training Details

In this section, we introduce the details of our training scheme for human and animal facial models.

Human At the preprocessing steps, we use Graphonomy [6] for selecting the interest region. Among 20 labels in Graphonomy, the label 'hair', 'face', and 'sunglasses' is contained in the interest region. We include the label 'sunglasses' for the following two reasons: First, a considerable number of images in the training dataset, which is used in training StyleGAN, contains faces with glasses on. Thus, StyleGAN is capable of generating various distributions of glasses. Second, in most cases, glasses are overlapped with the interest region, *i.e.*, facial region. Consequently, excluding the region for glasses occurs undesirable artifacts on the interest region for inverting real-world images with glasses. For the dilation, we empirically set the size of Gaussian Kernel to be (50,50) to contain the boundary information. We show examples of preprocessed images in Figure 1. We set L_{image} as below:

$$L_{image} = 1 \cdot L_2 + 0.8 \cdot \text{LPIPS} + 0.1 \cdot L_{ID}$$

Animal We use DEtection Transformer [4] for selecting the interest region. We use every part which consists of the animal as the interest region. In the animal domain, we use L_{MoCo} instead of L_{ID}, which is based on the MoCo v2 [8] network. Since the image resolution of the animal dataset is half of the human dataset, we set the size of Gaussian Kernel to be (25,25) for the dilation, which is a half size compared to FFHQ. We set L_{image} as below:

$$L_{image} = 1 \cdot L_2 + 0.8 \cdot \text{LPIPS} + 0.5 \cdot L_{MoCo}$$

We use Ranger [19] optimizer with learning rate 10^{-4}, and train the model for 500,000 steps with batch size 4. We save the model per every 10,000 steps, and lastly, select the model with the lowest validation L_{image}. We train on the single V100 GPU in an NVIDIA DGX1 machine, which takes around two weeks

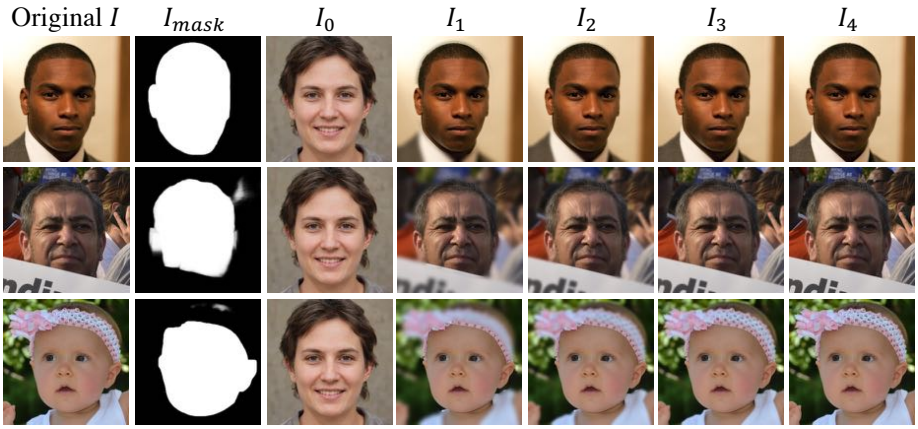

Fig. 1: **Preprocessing.** Preprocessed images using uninterest filter for training. In the uninterest region, we blur severely at the beginning of the iteration, as shown in I_1 and I_2. Gradually, images get clear, and we use the image at the last iteration, I_4, which becomes the same with I.

for the training. We set $N = 5$, the same as Restyle [2].[3] In UnF, we applied Gaussian blur with $r_i = (N - i)^2$, which was empirically set to fit our purpose.

B Dataset description

In this section, we introduce the dataset FFHQ [9], CelebA-HQ [13], and AFHQ [5], which are used for training and testing, among IntereStyle and other baseline models.

FFHQ Flickr-Faces-HQ (FFHQ) dataset has been released by StyleGAN [9], consisting of 70,000 high-quality facial images with the resolution 1024^2. Images are automatically aligned with Face Alignment Network(FAN) [11, 10] using landmark information. FFHQ contains only images under permissive licenses. We did **not** include any additional images except FFHQ for training IntereStyle.

CelebA-HQ CelebA-HQ dataset contains 30,000 high-quality facial images with resolution 1024^2. In the original CelebA dataset, splits for train-val-test are labeled, which we directly used for selecting the test dataset of CelebA-HQ dataset. Around 2,800 images are used for the test dataset of CelebA-HQ. All of the quantitative results are driven from the test dataset of CelebA-HQ.

[3] We maintain the training method of Restyle and *pSp* [15] as much as possible.

AFHQ AFHQ contains three domains of animal images: dogs, cats, and wild, with resolution 512^2. In this paper, we only used wild domain datasets for training and validation. We followed the split of the original dataset, which consists of 4,730 training and 500 validation images.

C Baseline Models Description

In this section, we briefly explain the baseline models of StyleGAN inversion. For a fair comparison, we exclude models which tune the generator, such as Pivotal-Tuning Inversion [16] or HyperStyle [3].

Image2StyleGAN Image2StyleGAN [1] (I2S) directly optimizes the style latent, to lower distortion between a generated image and the target image. Since we directly optimize the latent, we can invert images without any training stages. However, we need to optimize latents per every inversion, which is extremely time-consuming. Despite the prominent performance and simplicity, I2S is rarely used because of its excessively long inference time. Moreover, as shown in Figure 6 in the main paper, I2S causes artifacts in the interest region.

In-Domain GAN Inversion In-Domain GAN inversion (IDGI) [21] targets to learn the latent space of GAN for reconstructing the input image by decoding the latent space. IDGI uses a *domain-guided encoder*, which directly reconstructs the image on the real-world image space rather than the latent space. Moreover, *domain-regularized optimization* enables avoiding the stuck in the local minimum and uses a domain-guided encoder as a regularizer for preventing an out-of-domain inversion. For a fair comparison, we only use the domain-guided encoder for inversion without additional optimization steps per image.

pSp pixel2Style2pixel (*pSp*) [15] proposes a simple pyramid[12] structure for StyleGAN inversion, which effectively lowers distortion without any additional optimization steps. For utilizing the disentangled property of StyleGAN latent, *pSp* extracts coarse and fine features separately at the high and low layers, respectively. *pSp* provides various applications of StyleGAN inversion, such as image inpainting[20] and face frontalization[7].

e4e encoder4editing (*e4e*) [18] points the trade-off between distortion-perception is related to the distance of latent from W space. Consequently, for maintaining the latent close to W, *e4e* minimizes the variation of latents per layer and uses the latent discriminator at the training stage. Though *e4e* shows higher distortion than *pSp*, it shows high perceptual quality and editability simultaneously.

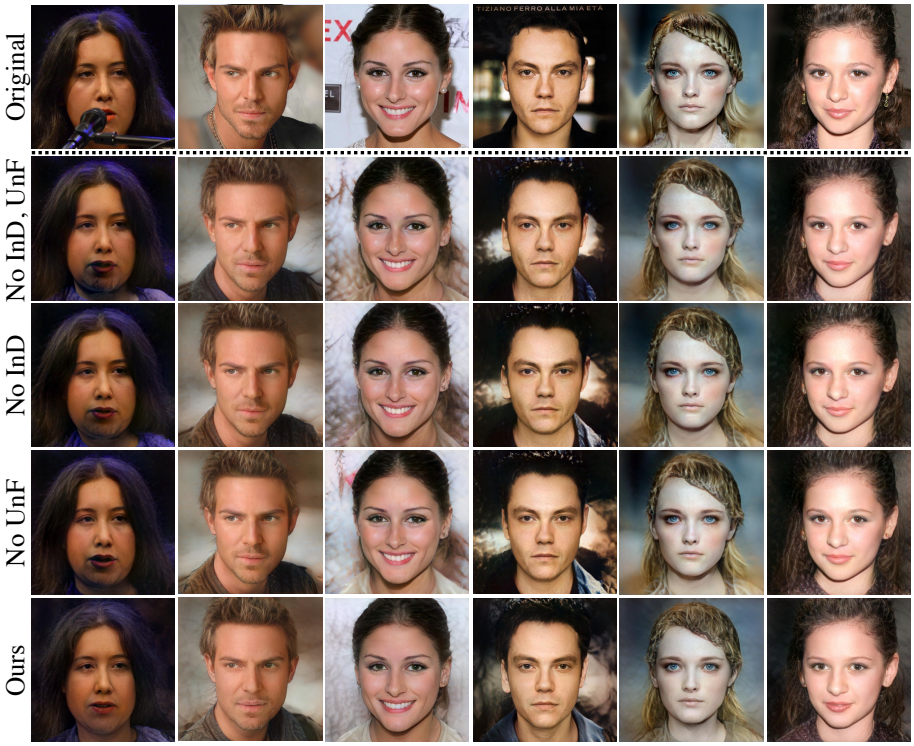

Fig. 2: **Ablation study for inversion.** Inversion results of various ablation models, with ours. Ours ignores noisy backgrounds and focuses on the interest region. Consequently, ours can invert the interest region robustly even with obstacles or blurs (first and second columns, respectively), while ablation models cannot. Since ours focuses on the interest region, it blurs the noisy backgrounds, as shown in the third and fourth columns. Moreover, images from our model show higher perceptual quality than baselines, which can be best viewed in zoom-in.

Restyle Restyle [2] claims that the single forward inference is not enough to utilize every information in the real-world image. Thus, Restyle uses iterative refinement, which gradually encodes the difference between a pair of input images. Starting from the image from the average latent of StyleGAN, Restyle step-wisely updates the image to approach the target image. Restyle achieves the lowest distortion at the time. All of the results are derived from iterating five steps, as the original Restyle did.

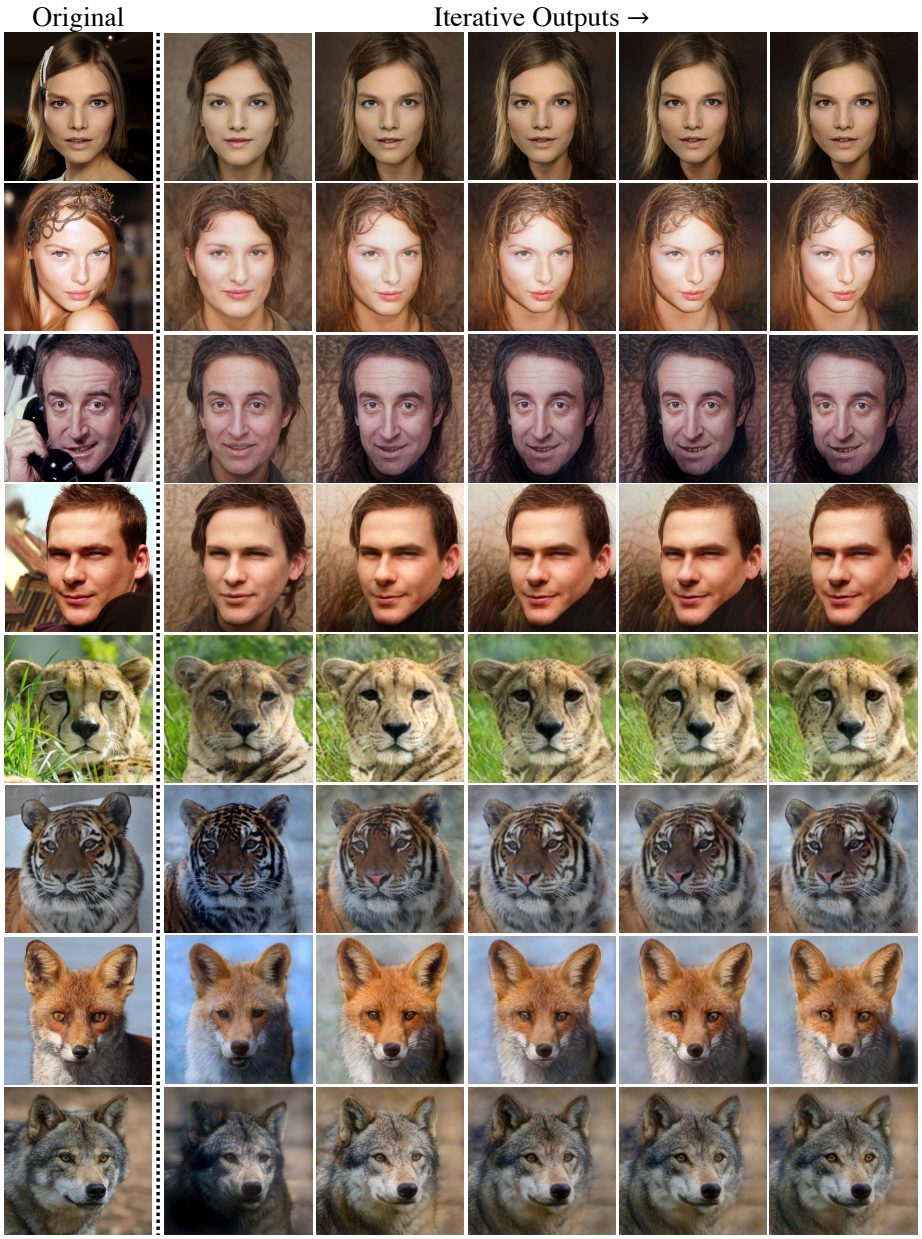

Fig. 3: **Various inversion results.** Various inversion results of our model on both human and animal datasets. We plotted the inversion images along with the intermediate outputs during iterative iteration. For every image, we progressed five iterations, same as the original Restyle. All images are sampled from CelebA-HQ or AFHQ test dataset.

D Qualitative Results

D.1 Inversion Results

In Figure 2, we compare inversion results with various ablation cases, qualitatively. We exclude Interest Disentanglement(InD) and Uninterest Filter(UnF) one by one, which yields three ablation models. For the clean image (leftmost in Figure 2), the quality marginally differs per each model. In the case when obstacles overlapped on the interest region (first column in Figure 2), simply applying masked loss without InD, and UnF is not enough to remove the artifacts. InD and UnF both are helpful for removing the artifacts, and artifacts are minimized when both methods are applied together. Moreover, our model shows robust inversion even within deformation, *i.e.*, blur on the interest region during alignment, as shown in the second column. When images have noisy backgrounds (third and fourth column in Figure 2), UnF simplifies the background noise, which makes the model focus on the interest region, and makes images more realistic. Finally, every image shows better perceptual quality with our model, which can be viewed in zoom-in.

We show more inversion results of our model, along with intermediate outputs during the iterative refinement process, in Figure 3. Our model starts from the image relatively close to I_0 and gradually updates it to be closer to I.

D.2 Editing via Latent Manipulations

In this section, we show additional results for image editing via latent manipulation methods, such as InterFaceGAN [17], and StyleCLIP [14].

First, Figure 4 shows the comparison of editing results of our model with various ablation cases, through InterFaceGAN. We compared editing results with a clean image (left), and an image where the obstacle overlaps the interest region (right). For the clean image, InD dramatically reduces undesirable artifacts, *e.g.*, eyeglass in the smile, and increases editability (See the fourth column of Figure 4). For the obstacle cases, InD and UnF mitigate the artifacts, in both inversion and editing cases. Moreover, InD shows a synergistic effect when combined with UnF, *i.e.*, ours, which increased editability for editing an age and mostly reduced the artifacts for both inversion and editing cases of the image with obstacles.

Figure 5 shows additional editing results of our model with StyleCLIP. To prove the high editability, we did experiments on both human and animal domains. In human cases, we showed the results for editing with two prompts: "curly hair" and "orange". Not to mention the precise inversion results, our model showed high editability with both prompts. Though the prompt "orange" is not giving an intuitive editing direction for human faces, our model reflected the prompt without harming the perceptual quality of given images. Our model showed high editability with prompts "lion" and "white", along with robust inversion results in animal cases. In the animal domain, several animals have large domain gaps, *e.g.*, wolves and lions [14], which is harsh to edit between two domains. However, our model showed high editability between animals from

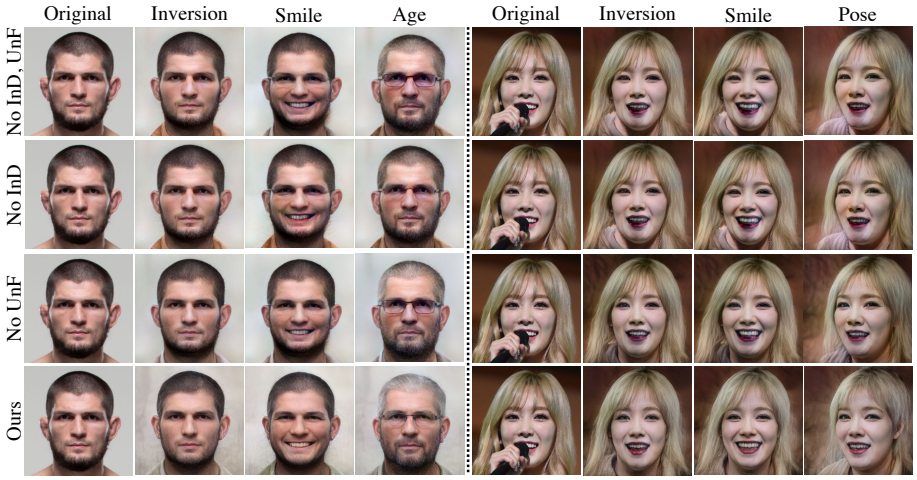

Fig. 4: **Ablation study for image editing.** Image editing results of various ablation models, with ours. We compare in two cases: a clean image (left) and a image with an obstacle (right). Even in the clean image, images show low editability without InD and UnF. Baseline models makes eyeglass artifacts for smile, and fails to aging the face in the desirable way. For the case of an obstacle, not only inversion but also edited images contain artifacts around the obstacle, which is significantly mitigated by our model.

the distance domain. Moreover, with the ambiguous prompt "white", our model reflected the prompt without harming the perceptual quality of the original images.

Figure D.2 showed additional results for style mixing [9]. We totally *randomly* sampled images from CelebA-HQ test dataset, to measure the consistency of the performance of our model, in various cases. Even several images contain difficult features for inversion, *e.g.*, overlapped hands and shoulders, extreme pupil direction, our model showed consistently high perceptual quality on style mixing. We want to note that GAN inversion models which tune a generator, *e.g.*, Pivotal Tuning [16], and HyperStyle [3], undergo several difficulties for style mixing, due to the difficulties of representing several images into a latent space of a single generator.

E Potential Negative Societal Impacts

Our research offers an inversion model of StyleGAN, which enables several robust editing without a shift of the identity. Consequently, one can maliciously manipulate the face of the random person simply by using a single photo without permission. Our robust inversion and editing results make the synthesized image be hard from distinguished clearly. We can partially prevent the abuse of

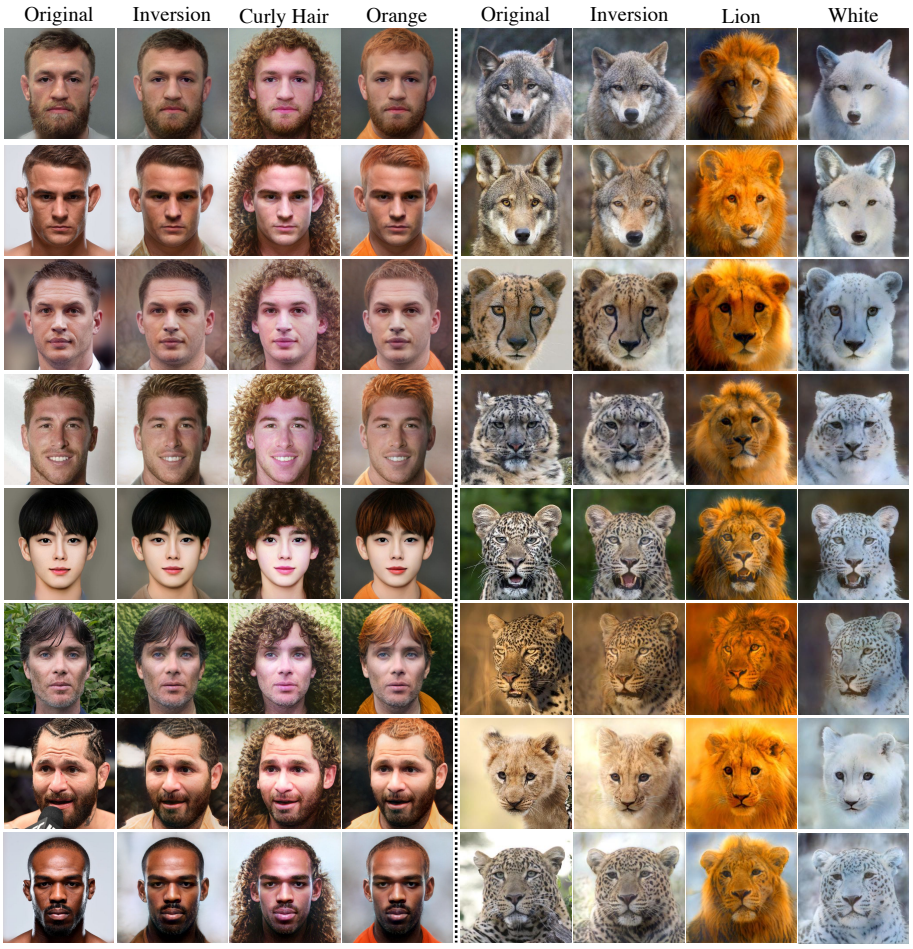

Fig. 5: **Various editing results.** Editing images via StyleCLIP, in human and animal domain. Our model showed consistently robust results on both inversion and editing scenarios. In both domain, our model showed robust editing not only for the explicit prompt, *e.g.*, "curly hair" and "lion", but also for the implicit prompt, *e.g.*, "orange" and "white".

our model by restrictively releasing the pre-trained model, *e.g.*, as a format of API, or do not release the editing boundaries that can be used maliciously, *e.g.*, editing boundaries related to the sexual or racism-related facial expression.

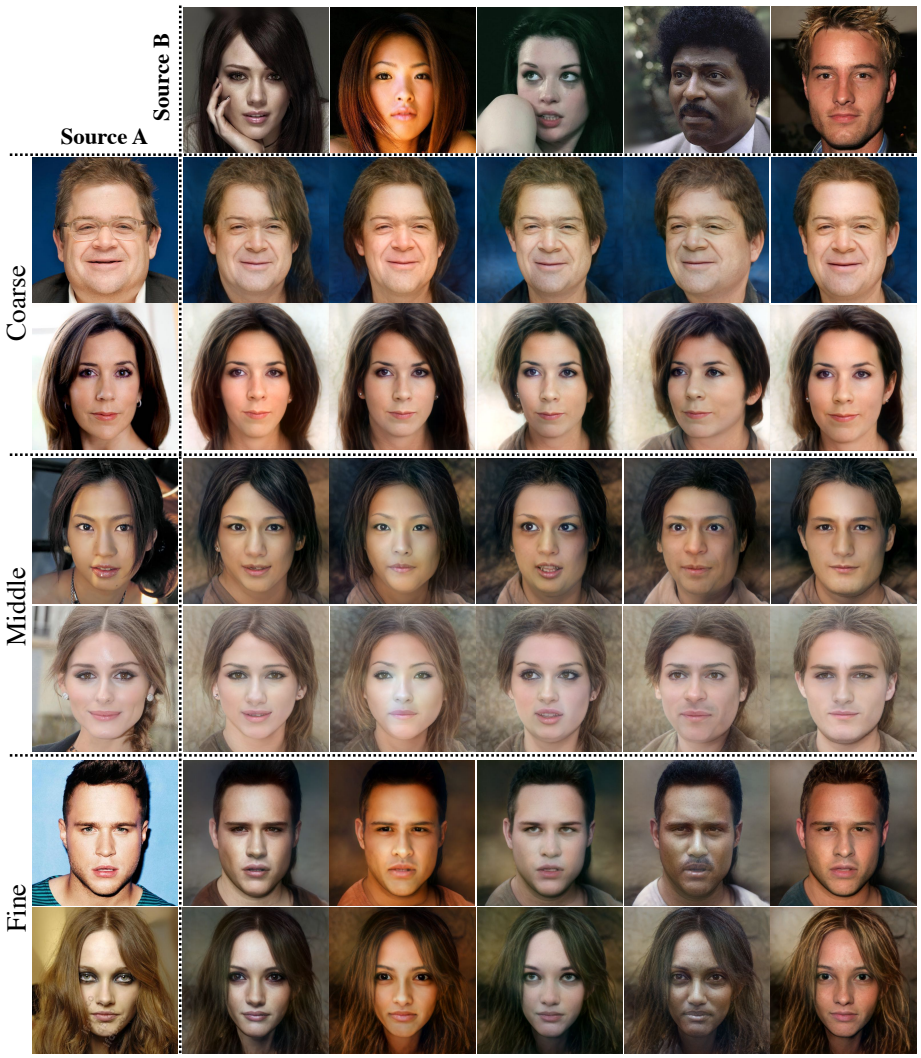

Fig. 6: **Various style mixing results.** Style mixing results of images *randomly* selected from CelebA-HQ test dataset, by our model. Following the settings from the original StyleGAN [9], we extract coarse, middle, and fine styles from source A, in order, and the rest style from source B. Though several images contain features that can yield artifacts, *e.g.*, overlapped hands or shoulders, irregular lights, our model showed robust style mixing results consistently.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4432–4441 (2019)
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. Alaluf, Y., Tov, O., Mokady, R., Gal, R., Bermano, A.H.: Hyperstyle: Stylegan inversion with hypernetworks for real image editing. arXiv preprint arXiv:2111.15666 (2021)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
5. 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)
6. Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., Lin, L.: Graphonomy: Universal human parsing via graph transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7450–7459 (2019)
7. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4295–4304 (2015)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
9. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
10. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1867–1874 (2014)
11. Kowalski, M., Naruniec, J., Trzcinski, T.: Deep alignment network: A convolutional neural network for robust face alignment. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 88–97 (2017)
12. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
13. Liu, Z., Luo, P., Wang, X., Tang, X.: Large-scale celebfaces attributes (celeba) dataset. Retrieved August **15**(2018), 11 (2018)
14. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021)
15. 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)
16. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. arXiv preprint arXiv:2106.05744 (2021)

17. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE transactions on pattern analysis and machine intelligence (2020)
18. 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)
19. Yong, H., Huang, J., Hua, X., Zhang, L.: Gradient centralization: A new optimization technique for deep neural networks. In: European Conference on Computer Vision. pp. 635–652. Springer (2020)
20. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018)
21. 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)