Discovering Interpretable Latent Space Directions of GANs Beyond Binary Attributes — Supplementary Materials —

Huiting Yang[1], Liangyu Chai[1], Qiang Wen[1], Shuang Zhao[2], Zixun Sun[2], Shengfeng He[1*]
[1] School of Computer Science and Engineering, South China University of Technology
[2] Interactive Entertainment Group, Tencent Inc.

1. Introduction

In this document, we provide additional experiments to further examine our method, AdvStyle. In the first section, we provide more implementation details. In the second section, we elaborate how we collect and process our training datasets. Lastly, we show additional experiments results including the limitation of our proposed method, statistical measurement, editing results and analysis of binary and non-binary attributes, and real image manipulation comparison.

2. Implementation details

2.1. Distribution of Training Samples

The latent code z is sampled from the Gaussian distribution $\mathcal{N}(0, I_d)$, where $d = 512$ is the dimensionality of the latent code. The selected index k is sampled from a uniform distribution $\mathcal{U}\{1, K\}$, where $K = 100$. The step size α is sampled from a uniform distribution $\mathcal{U}\{-6, 6\}$.

2.2. Comparison Details

When comparing to InterFaceGAN [12], five human face attributes are provided in their released codes, including *old, smile, pose, eyeglasses*, and *female*, we directly use these well-trained directions. The other human style attributes and anime attributes are trained based on their provided model and default parameters. These attribute assessors obtain higher than 95% accuracy on the validation set, which is higher than the accuracy reported in their paper [12].

When comparing to GANSpace [5], four human face attributes are provided in their released codes, including *smile, pose, eyeglasses*, and *female*, we directly use these well-trained directions.

*Corresponding author (hesfe@scut.edu.cn). This work is supported by the National Natural Science Foundation of China (No. 61972162), and CCF-Tencent Open Research fund. Code is available at https://github.com/BERYLSHEEP/AdvStyle.

2.3. Pretrained Generator Models

For animate attributes editing, the generator of StyleGAN [7] is trained on the Danbooru2018 dataset [4]. For human face attribute editing, the generator of StyleGAN is trained on the FFHQ dataset [7]. Note that all the results are generated by these generators, *i.e.*, all images share the same learned latent space, and the only difference is how semantic directions are discovered. Some images may show water droplet -like artifacts, which is a defect of the StyleGAN network [8].

3. Attribute Datasets

For animate attributes, we aim to generate anime characters with a high-resolution of 512×512. Some publicly available anime datasets do not meet our needs, *e.g.*, the images in [6] are with a low resolution of 64×64. To obtain datasets with diverse attribute labels, we collect 9 attribute datasets, corresponding to 6 character attributes and 3 anime styles. All the images are resized to 512×512 for training.

1. Danbooru2018 dataset [4] provides large-scale high-resolution anime images associated with metadata. However, the original images of the dataset contain multiple characters in the same scene. We detect the anime faces using a faster-RCNN [11] based anime face detector [1], then crop the image at a larger scale of 1.2 to include the hair of the character. We use the above preprocessing method to obtain 7 attribute datasets, including 6 character attributes of *open mouth, blunt bangs, short hair, black hair, blonde hair, pink hair* and 1 style attribute *Itomugi-Kun*. Each attribute has more than 800 images.

2. Manga 109 [10, 3] consists of 109 comic books of 21142 pages drawn by professional artists. We extract 2689 anime faces according to the annotations with a

[1] https://github.com/qhgz2013/anime-face-detector

Figure 1: A failure example on *realness* attribute.

similar preprocessing above. These images construct a *comic* style attribute dataset.

3. To simulate the real anime style of a famous artist, we manually collect an anime style dataset from Chibi Maruko-chan. It contains around 1000 images.

For human face attribute editing, five binary attributes (*pose, old, female, smile, eyeglasses*) are trained using the CelebA dataset [9]. Although CelebA dataset contains binary attribute annotations, we only use the positive samples for binary attributes, *e.g.*, only the *old* attribute is collected and the negative *young* direction is learned in an unsupervised manner. Two style attributes (*supermodel style* and *Chinese celebrity style*) are trained on datasets collected from [1] with 1024×1024 resolution.

4. More Results

4.1. Failure Case

Latent space exploration methods cannot produce results that exceed the generation ability of the pre-trained generator. Therefore, even though AdvStyle can learn from any positive samples, the transformed results may not be satisfactory for those attributes are too far from the original domain. We explore this limitation by learning a *realness* attribute to convert anime characters to human using the FFHQ training dataset. As shown in Fig. 1, we cannot map an anime character to a human face, but learn a direction to simulate real human face distributions, like adding face-shading effects.

4.2. Statistical Measurement

As shown in Fig. 2, we utilize the FID scores to measure the diversity and quality of the editing results. Compared to InterFaceGAN, the FID scores of AdvStyle are more stable and closer to the original scores in both binary and non-binary attributes. Even for the challenging multi-attributes manipulation, the FID scores only slightly degrade which shows the learned directions are disentangled to each other.

4.3. Real Image Manipulation

Here we supplement more real image manipulation results by comparing with three methods, StyleRig [13], StyleFlow [2], and InterFaceGAN [12]. We have not compared with StyleRig and StyleFlow in the main text for the

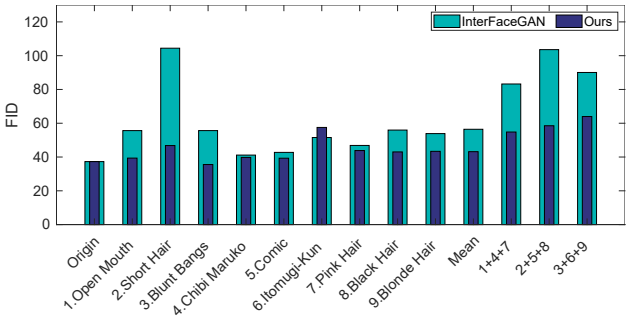

Figure 2: FID scores for single and multiple attributes manipulation with anime characters.

Figure 3: Real image manipulation.

following reasons. First is that StyleRig uses 3D vertices for training (as discussed in Table 1 of the main text), which is substantially different from image-only training setting. Second is that StyleFlow was not a formal publication at the date of the CVPR deadline.

Despite all that, we have further compared with these methods in Fig.3. As StyleRig does not release the source code, we directly copy the images from the paper, and apply the official implementation of StyleFlow to generate the outputs. We can see that, although StyleRig is jointly trained with *pose, smile* and *illumination* attributes, both the *pose* and *smile* attributes are entangled with illumination. For StyleFlow, the results are with blurry artifacts, especially for the *smile* attribute. The editing results of InterFaceGAN create unnatural artifacts on *pose* attribute. Furthermore, we challenge the failure case (the right example of Fig. 3) presented in StyleRig, and the editing results show that our method achieves better disentanglement and identity preservation.

4.4. Attribute Manipulation Results

From Fig. 4 to Fig. 19, these results supplement the editing results in the main submission and here we will focus on attributes that are not discussed in the paper.

Binary Attributes Manipulation. For human face at-

tribute editing, we mainly compare AdvStyle with the supervised binary method InterFaceGAN [12] and unsupervised method GANSpace [5], and the latter supports four attributes of *smile, pose, eyeglasses,* and *female*. Although unsupervised method can discover unexpected attributes, the found directions are not disentangled as supervised method does. As shown in the bottom right corner of Fig. 4, the headband disappears when editing the *smile* attribute with the direction learned by GANSpace. We also find that *pose* (Fig. 5) and *eyeglasses* (Fig. 6) directions learned by GANSpace are entangled with hairstyle and smile, respectively.

On the other hand, both supervised methods InterFaceGAN and our AdvStyle achieve similar editing results on the simple binary attributes like *smile, pose,* and *eyeglasses* (Fig. 4, Fig. 5, and Fig. 6). However, for binary attributes that involve global editing, like *old* and *female*, entanglement problem also occurs in InterFaceGAN. For example, the *old* direction is entangled with eyeglasses (Fig. 8), while the *female* direction is entangled with hairstyle, beard, and smile (Fig.7). Also, there are problems with the continuous attribute like *short hair* in Fig. 12, as the boundary of positive and negative samples are not distinct, leading to modification on irrelevant anime style. The comparison results show that our proposed method AdvStyle achieves better disentanglement.

Non-binary Attributes Manipulation. For non-binary attribute editing, the compared supervised method InterFaceGAN fails in most cases. As shown in Fig. 13, changing the *blunt bangs* direction of InterFaceGAN leads to large variations in anime style. Besides, the negative direction is uncorrelated with *bangs* attribute but produces an unknown hair style. For the *black hair* attribute in Fig. 14 and the *pink hair* attribute in Fig. 16, InterFaceGAN mainly finds a dark green hair direction and purple hair direction which are not correct. On the contrary, our AdvStyle can learn accurate directions not only in binary attributes but also in non-binary attributes. For example, the *comic* style strengthens the edges of the characters (Fig. 18) and the *maruko* style produce a baby-like character while preserving the original identity since the characters in Maruko dataset are cute with baby-like faces (Fig. 19).

References

[1] http://www.seeprettyface.com/. 2

[2] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv e-prints*, pages arXiv–2008, 2020. 2

[3] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. *IEEE MultiMedia*, 27(2):8–18, 2020. 1

[4] Anonymous, Danbooru community, and Gwern Branwen. Danbooru2019: A large-scale crowdsourced and tagged anime illustration dataset. https://www.gwern.net/Danbooru2019, January 2020. 1

[5] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *NeurIPS*, 2020. 1, 3

[6] Yanghua Jin, Jiakai Zhang, Minjun Li, Yingtao Tian, Huachun Zhu, and Zhihao Fang. Towards the automatic anime characters creation with generative adversarial networks. In *NeurIPS*, 2017. 1

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4396–4405, 2018. 1

[8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 1

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, December 2015. 2

[10] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 1

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99. 2015. 1

[12] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 1, 2, 3

[13] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *CVPR*, pages 6142–6151, 2020. 2

Figure 4: Manipulation on *smile* attributes. Images are generated by moving in positive or negative directions.

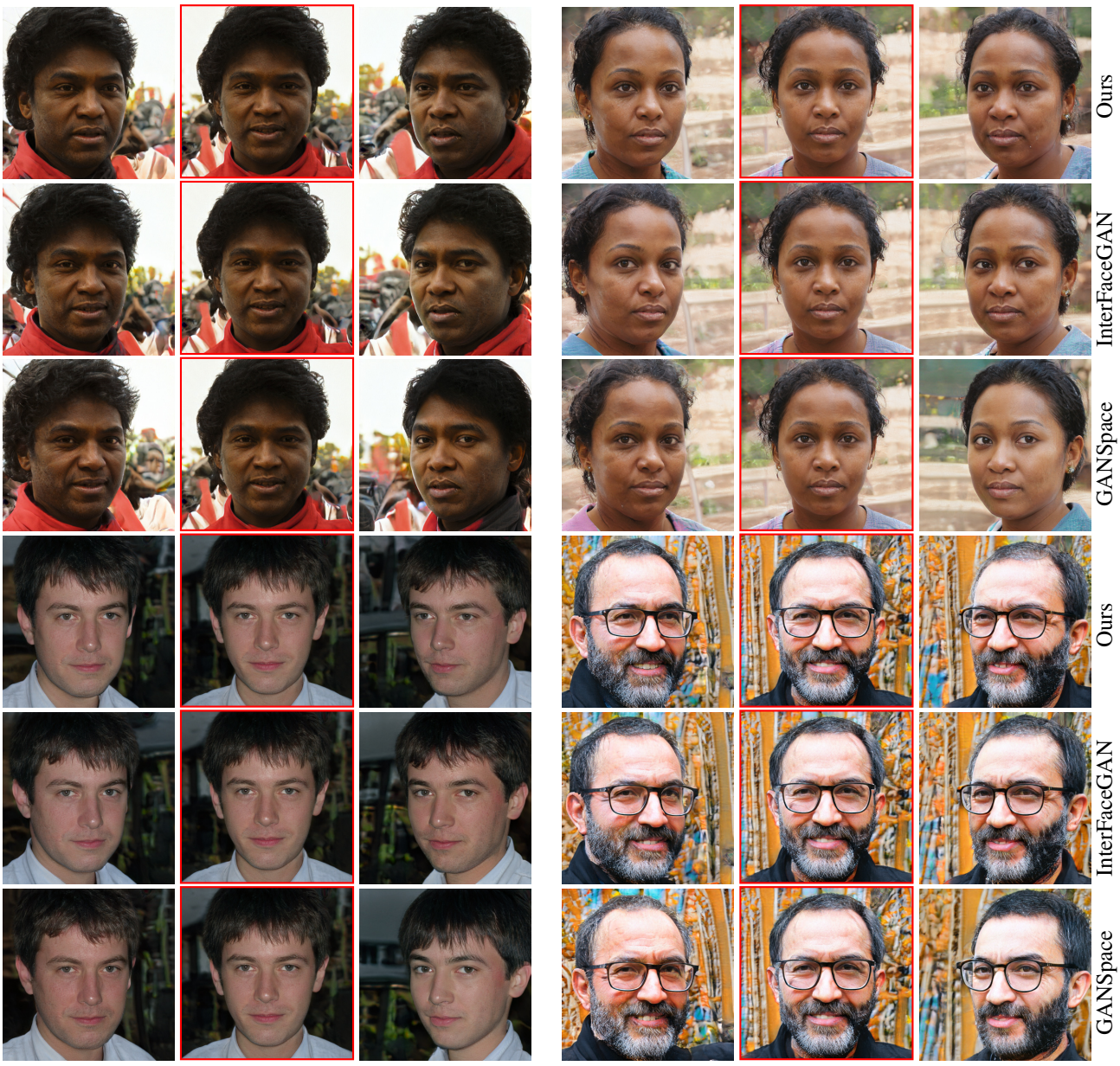

Figure 5: Manipulation on *pose* attributes. Images are generated by moving in positive or negative directions.

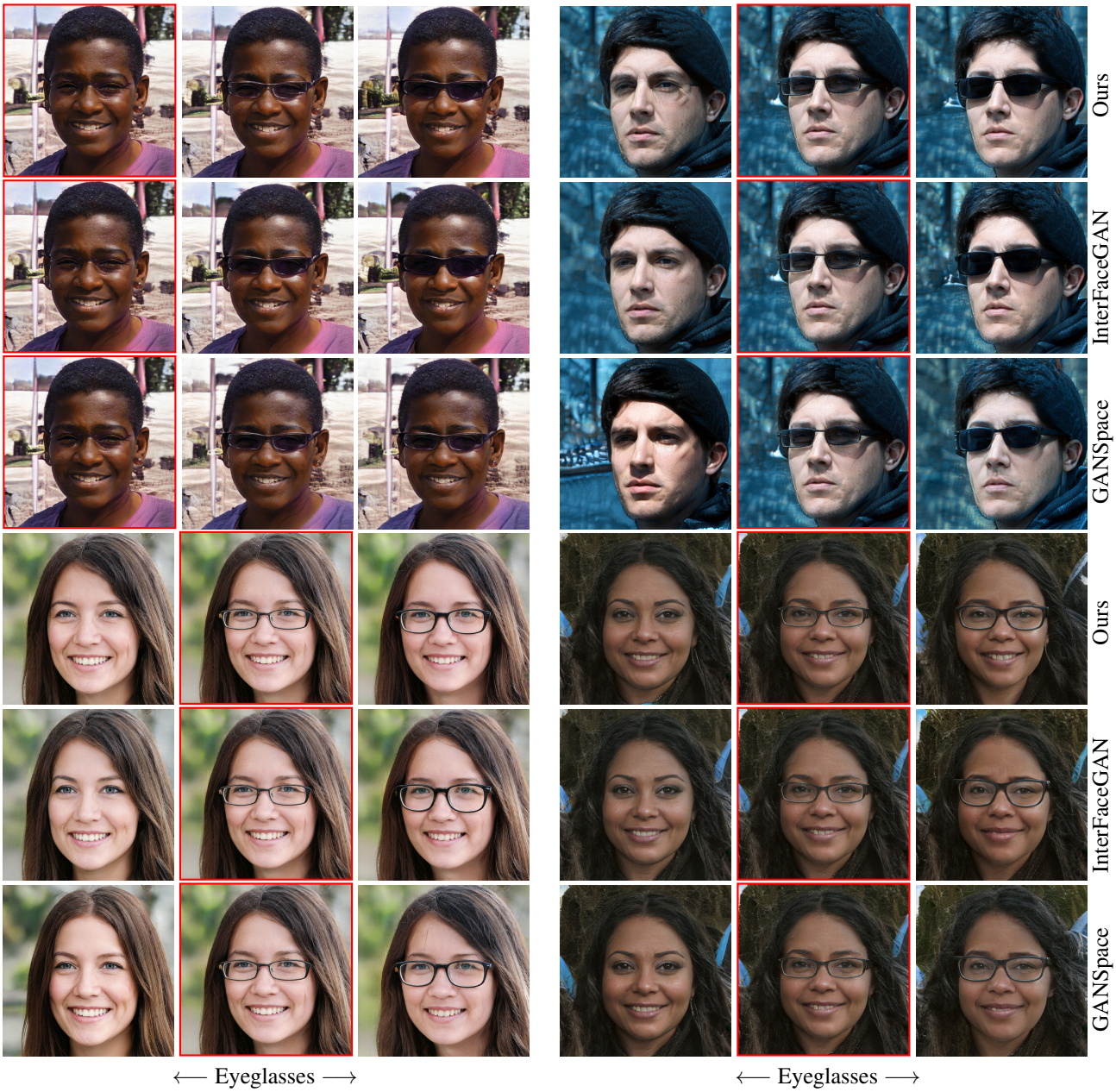

Figure 6: Manipulation on *eyeglasses* attributes. Images are generated by moving in positive or negative directions.

Figure 7: Manipulation on *female* attributes. Images are generated by moving in positive or negative directions.

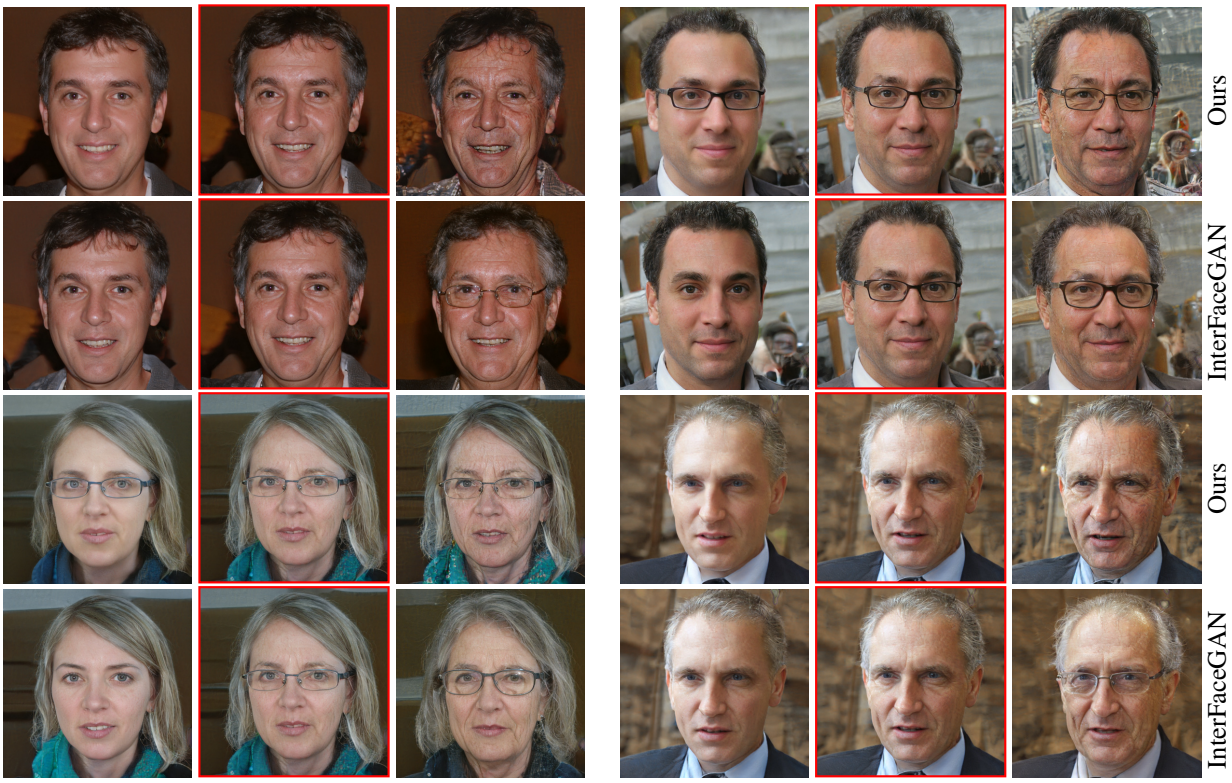

Figure 8: Manipulation on *old* attributes. Images are generated by moving in positive or negative directions.

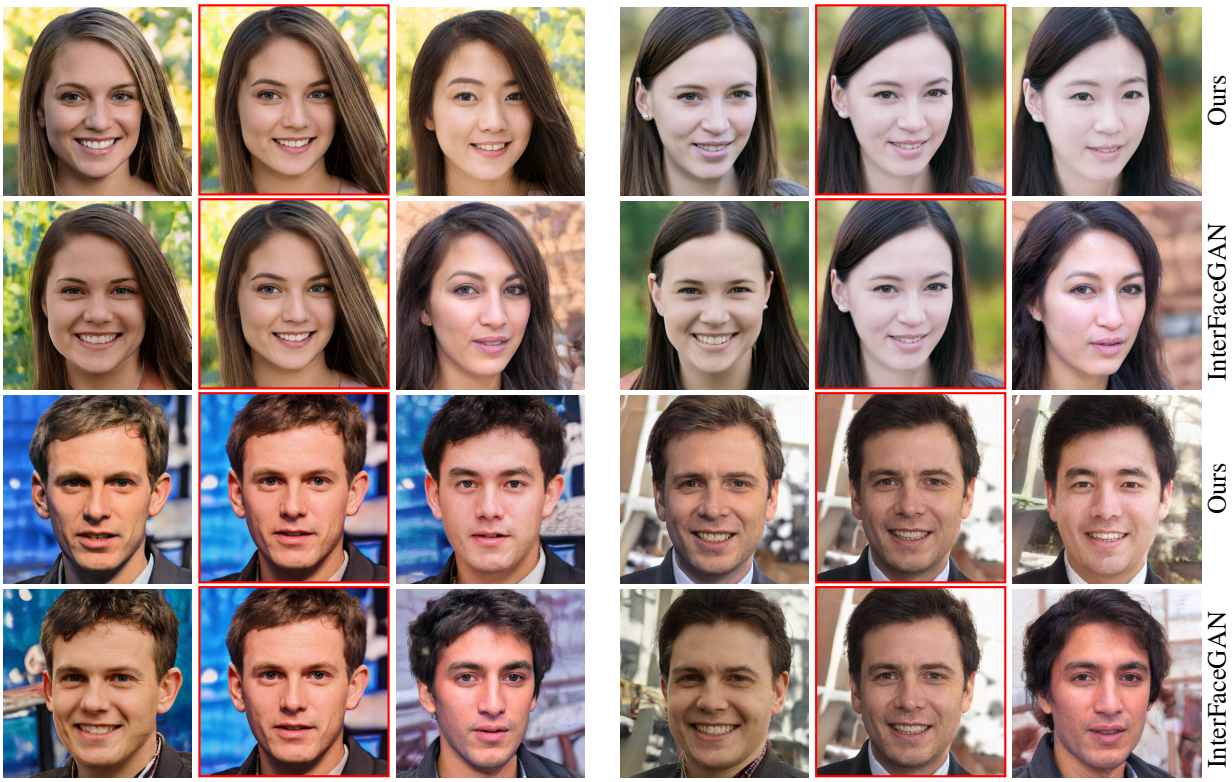

Figure 9: Manipulation on *Chinese celebrity* attributes. Images are generated by moving in positive or negative directions.

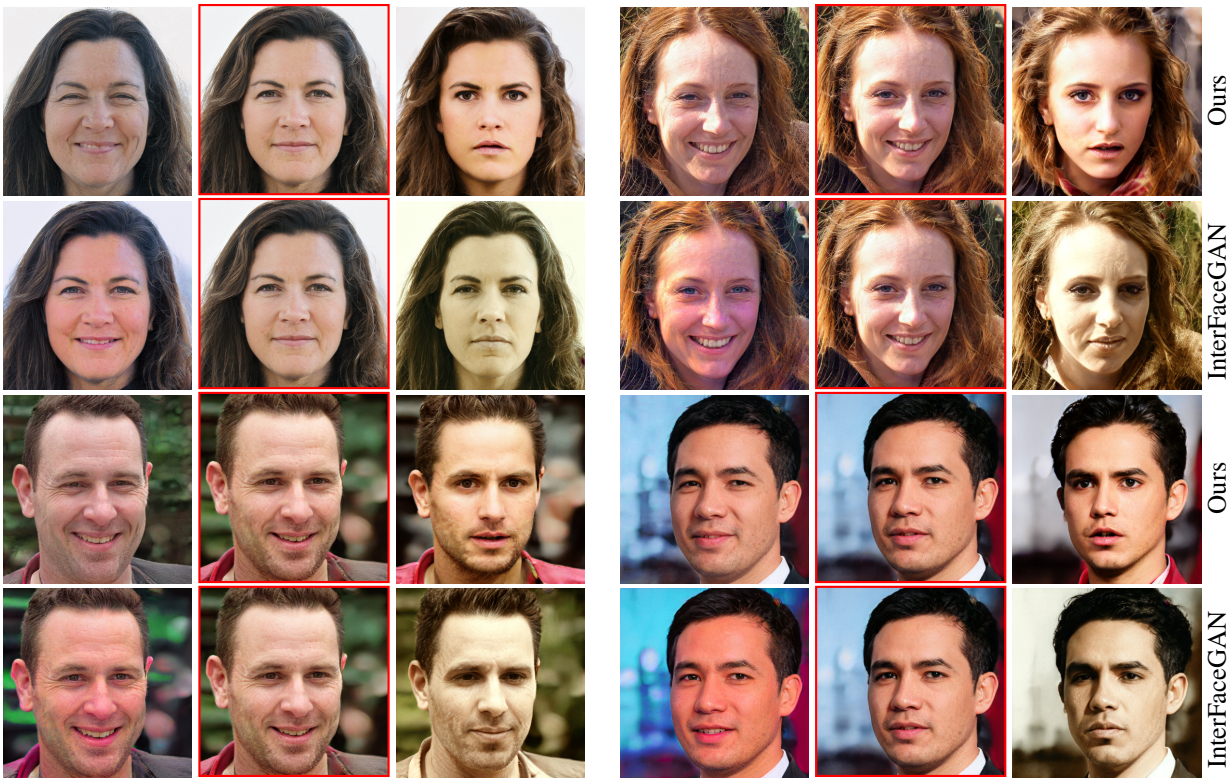

Figure 10: Manipulation on *supermodel* attributes. Images are generated by moving in positive or negative directions.

Figure 11: Manipulation on *open mouth* attributes. Images are generated by moving in positive or negative directions.

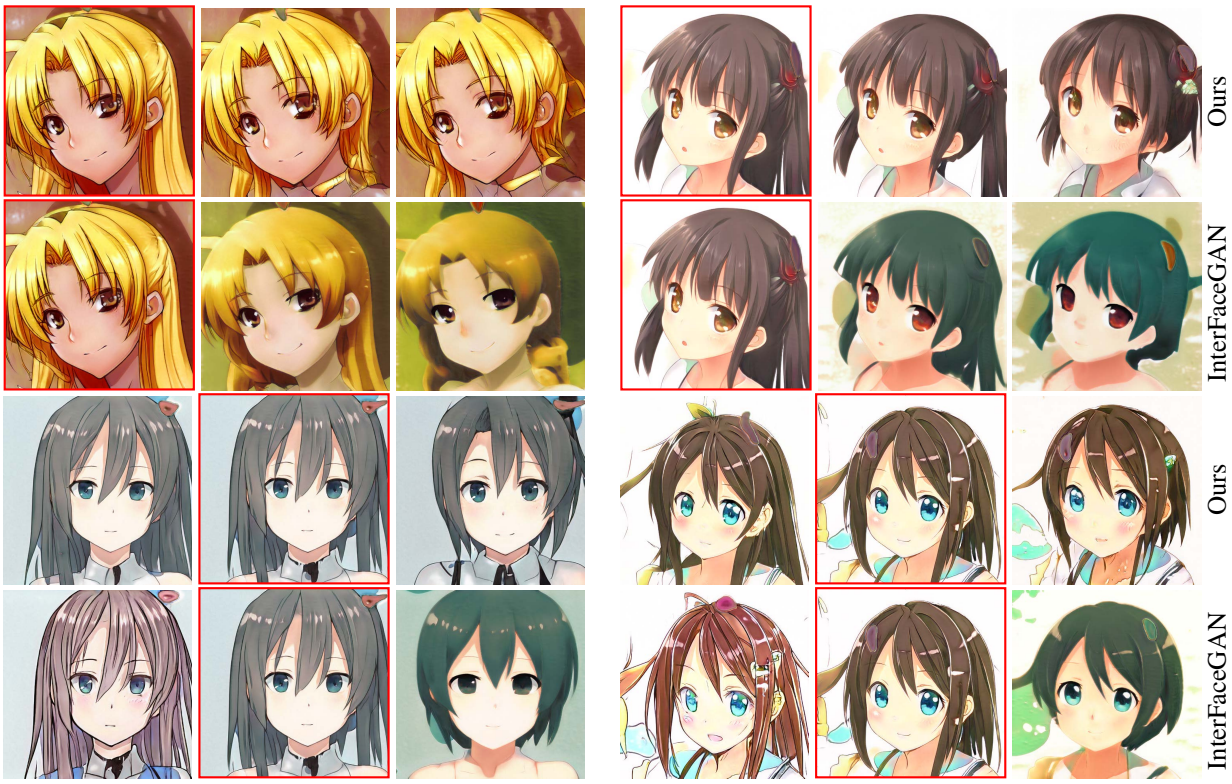

Figure 12: Manipulation on *short hair* attributes. Images are generated by moving in positive or negative directions.

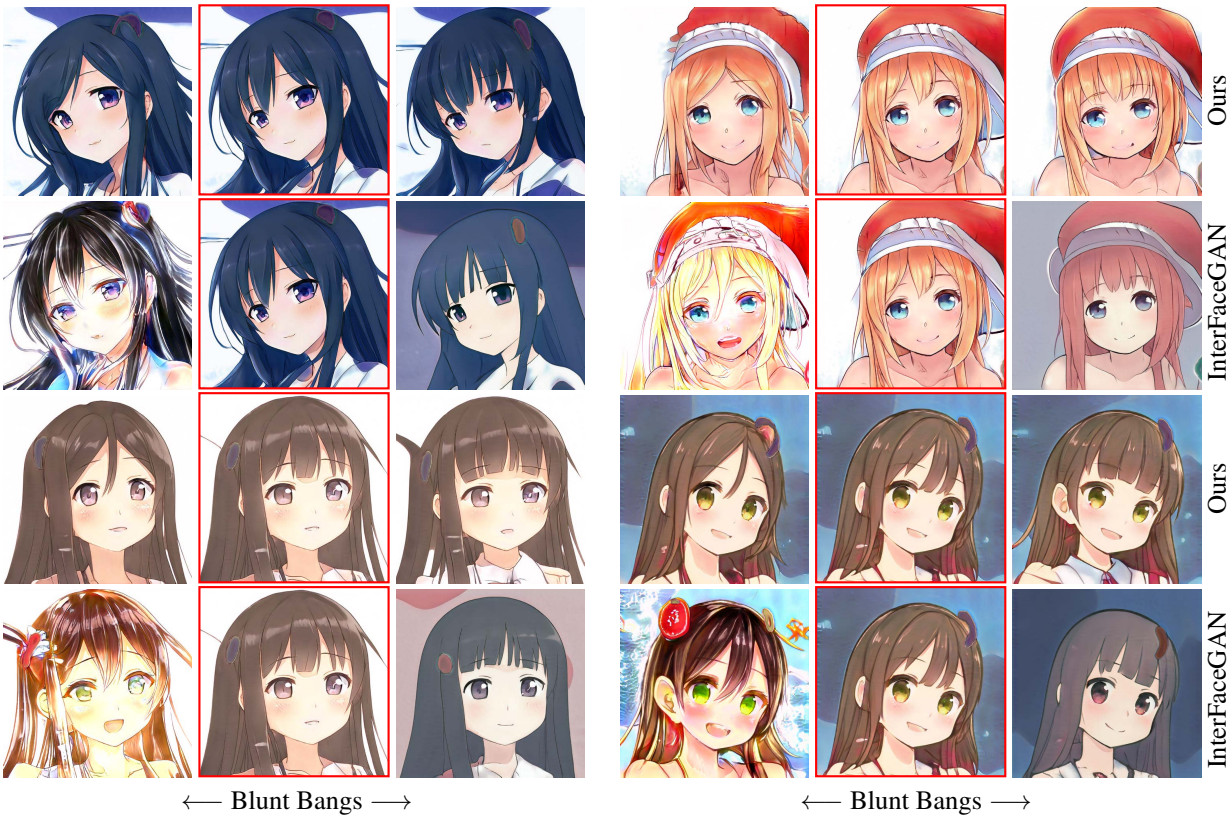

Figure 13: Manipulation on *blunt bangs* attributes. Images are generated by moving in positive or negative directions.

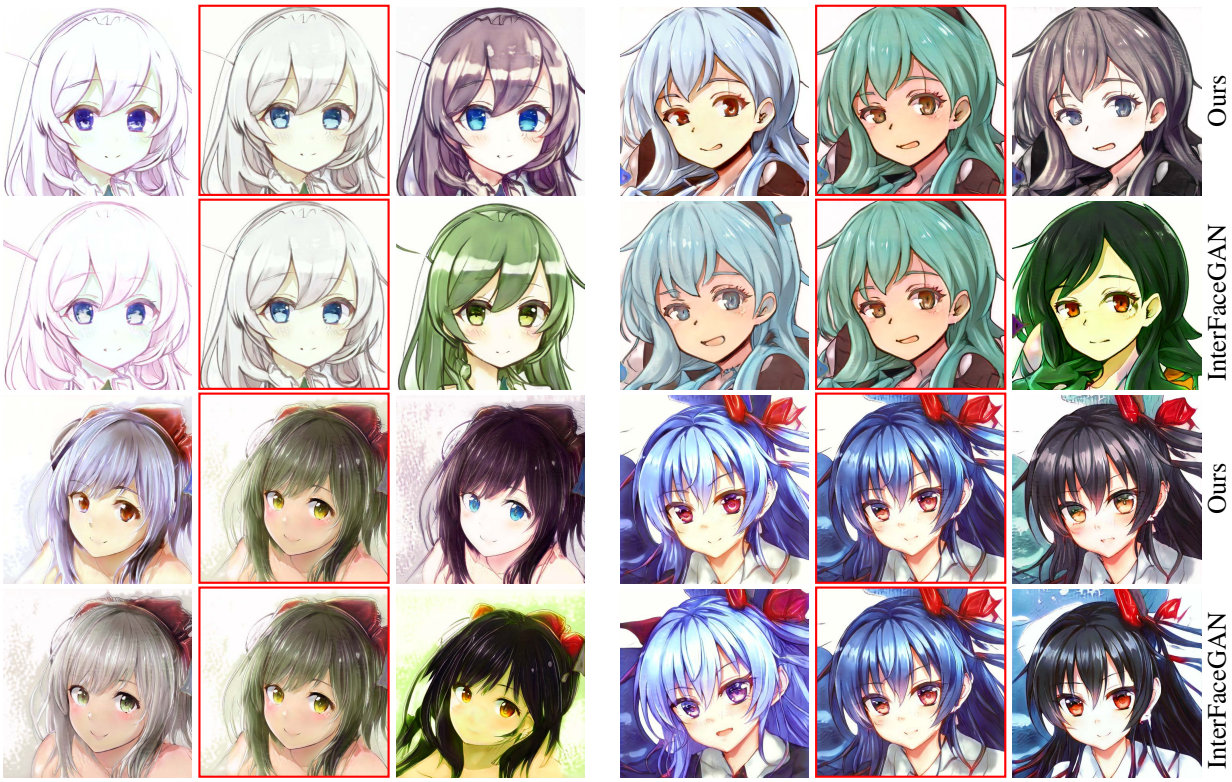

Figure 14: Manipulation on *black hair* attributes. Images are generated by moving in positive or negative directions.

Figure 15: Manipulation on *blonde hair* attributes. Images are generated by moving in positive or negative directions.

Figure 16: Manipulation on *pink hair* attributes. Images are generated by moving in positive or negative directions.

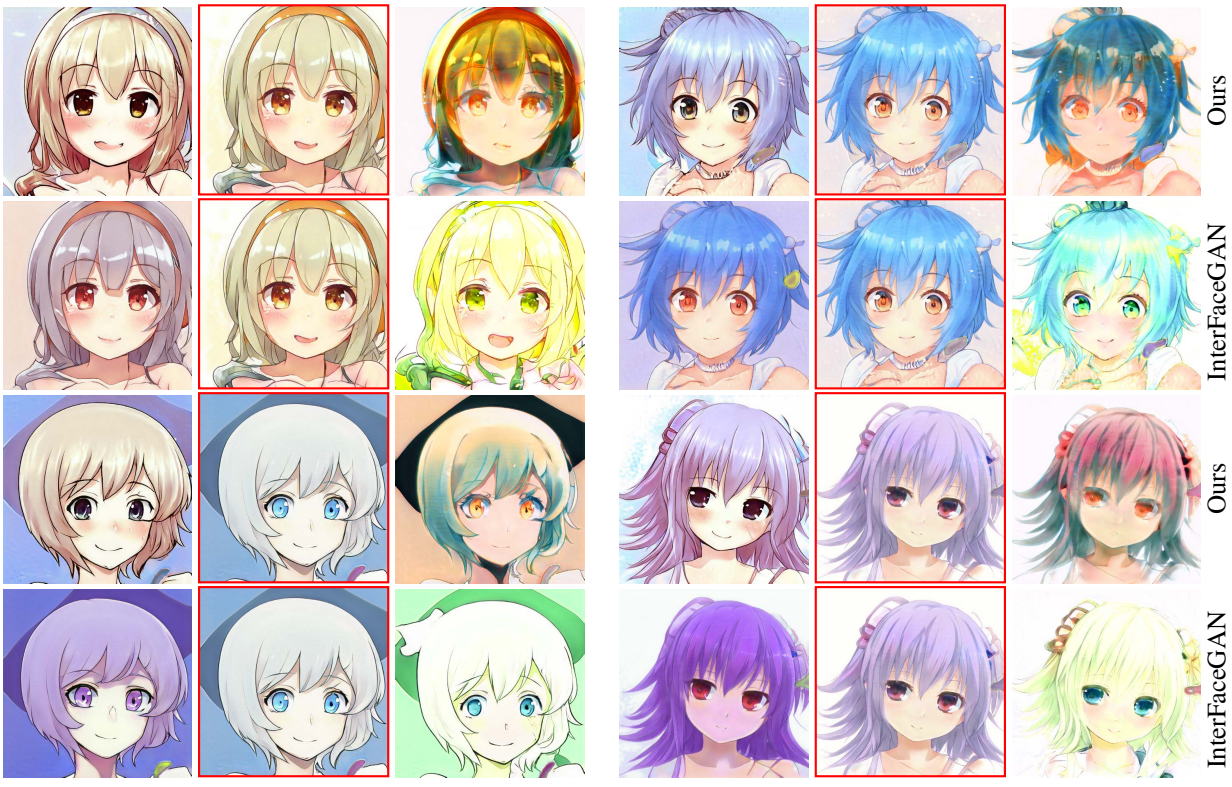

Figure 17: Manipulation on *Itomugi-Kun* attributes. Images are generated by moving in positive or negative directions.

Figure 18: Manipulation on *comic* attributes. Images are generated by moving in positive or negative directions.

Figure 19: Manipulation on *maruko* attributes. Images are generated by moving in positive or negative directions.