Supplementary material for Quantitative Manipulation of Custom Attributes on 3D-Aware Image Synthesis

Hoseok Do[1,2] EunKyung Yoo[1] Taehyeong Kim[3†] Chul Lee[1] Jin young Choi[2]

[1]AI Lab, CTO Division, LG Electronics, South Korea
[2]ASRI, Dept. of Electrical and Computer Engineering, Seoul National University, South Korea
[3]Dept. of Biosystems Engineering, Seoul National University, South Korea
{hoseok.do, eunkyung.ryu, clee.lee}@lge.com, {hoseok03, taehyeong.kim, jychoi}@snu.ac.kr

A. Potential Negative Societal Impacts

Applications with multi-view 3D-aware image synthesis and manipulation could be misused for criminal and misinformation. There are some fake detection researches for detecting whether the image is real or GAN-generated image [3, 14]. We strongly oppose the use of our methods for malicious purposes and hope that our method can be used to improve fake detection research and make positive scientific advances.

B. Implementation Details

B.1. Attribute Dataset Construction

In this section, we present implementation details for constructing attribute dataset $\mathcal{D}_a$. For attributes with public datasets, attribute datasets $\mathcal{D}_a$ were constructed by sampling from public datasets. For the face category's age attribute, the FFHQ-Aging [11] dataset was divided into ten groups ($M = 10$), which were obtained from 0–2 to 70 years old or older. For the face category's smile attribute, we used the KDEF-dyn [4] dataset, a video dataset that records changes in the facial expressions of 40 people. We captured 24 images from each video and divided them into eight groups ($M = 8$) in chronological order. Images for the car category's year attribute in the DVM-CAR [5] dataset were divided into 20 groups from 2000 to 2019 ($M = 20$). We used the Stanford Cars [9] dataset for the car category's type attribute. Five groups were used in the order of coupe, sedan, wagon, SUV, and minivan ($M = 5$). We used the DVM-Car [5] dataset for the car category's type attribute, and two sets of color attributes were tested: achromatic color and chromatic color. In the achromatic color experiment, four groups of black, gray, silver, and white were used ($M = 4$). In the chromatic color experiment, four groups of red, orange, yellow, and green were used ($M = 4$). For custom attributes that had no attribute dataset, we manually constructed $\mathcal{D}_a$ for each attribute. For the church category, darkness and cloud attributes were tested. For the darkness attribute, the images were divided into four groups ($M = 4$) midnight, twilight, dawn, and daytime; for the cloudiness attribute, the images were divided into five groups ($M = 5$) according to the cloud amount. For the custom attributes of the face and cat categories, we divided the groups according to the quantity of the attribute same as the manner used for the church category. The number of groups and the number of samples for each group is shown in Table 1.

B.2. Implementation on StyleGAN2

We trained the navigator for 20,000 iterations with a batch size of five for datasets with a resolution of 1024 $\times$ 1024 images (faces) and 10,000 iterations with a batch size of ten for datasets with lower-resolution images (cars and churches). The source latent feature w_s is encoded from the source image using the e4e encoder [13] during navigator training. Unlike EG3D, StyleGAN2 cannot generate multi-view images, so we did not use the view consistency factors in the experiments for StyleGAN2. Specifically, we did not use the consistency dataset and trained navigator without multi-view, which are the same settings as in the view consistency ablation study in Section 4.4.

[†]Work done during at LG Electronics

category	face								
attribute	age	smile	gender	eyeglasses	eyebrow†	pain†	thinness†	drunkenness†	tiredness†
M	10	8	2	2	2	2	5	3	2
Ω_a	100	10	100	100	50	20	20	10	50

category	cat			car				church	
attribute	age†	eye†	fur†	year	type	achro. color	chro. color	darkness†	cloudiness†
M	2	4	4	20	5	4	4	4	5
Ω_a	50	25	25	100	100	100	100	25	20

† indicates custom attributes implemented via manually constructed D_a.

Table 1. Number of groups (M) and the number of samples for each group (Ω_a) in constructing each attribute dataset $\mathcal{D}_a$.

B.3. Implementation of StyleFlow on EG3D

StyleFlow originally used Microsoft Face API [10] as a quantifier, but Microsoft no longer supports it. Therefore, we used Face++ [1] as a quantifier for StyleFlow implementation on EG3D. Among the attributes tested in StyleFlow, we tested with four attributes (age, smile, gender, and eyeglasses) which are supported by Face++. Like StyleFlow's original implementation on StyleGAN2, we randomly generated 10K latent features with truncation factor $\phi = 0.7$, then generated images from features. And we trained the flow model using the randomly generated 10K images and their attribute labels evaluated by Face++. The architecture and training settings are the same as StyleFlow's original implementation on StyleGAN2. For StyleFlow+Q, all other implementations are the same, except for the quantifier being replaced from Face++ with our custom attribute quantifier.

C. Additional Experiments

C.1. Color Jittering in Quantifier

In additional experiments, we evaluated the color bias in gender attribute manipulation (red for females, blue for males). We generated gender image sets, each consisting of a male image and a female image, by manipulating gender attributes from each source image. Then, we calculated the average background color bias due to gender manipulation. We calculated the average color difference value between the red channel image I^{red} and the blue channel image I^{blue}. Then, we calculated β from the difference between the average color difference values of male and female set images as follows:

$$\beta = (\bar{I}_F^{red} - \bar{I}_F^{blue}) - (\bar{I}_M^{red} - \bar{I}_M^{blue}), \tag{1}$$

where $\bar{I}$ denotes the average pixel intensity excluding one-quarter of the center of the image, and I_F and I_M are female and male images, respectively. Therefore, a higher β value means more bias; the image manipulated to the female image has a redder background than that manipulated to the male image. We compared our method to Ours-C, which did not use color-jittering augmentation while training the quantifier. As shown in Table 2, our quantifier contributes to declining the color bias in gender both in our navigator and StyleFlow.

	StyleFlow	StyleFlow+Q	Ours-C	Ours
$\beta \downarrow$	4.56	2.73	3.54	**2.56**

Table 2. Difference in average intensities of red and blue channels according to gender manipulation.

C.2. Manipulation on Real Image

For the attribute manipulation on real images, we used pivotal tuning inversion (PTI) [12] to obtain the latent feature from the real image. Figure 1 shows that various attributes can be manipulated on the latent feature inverted from the real image.

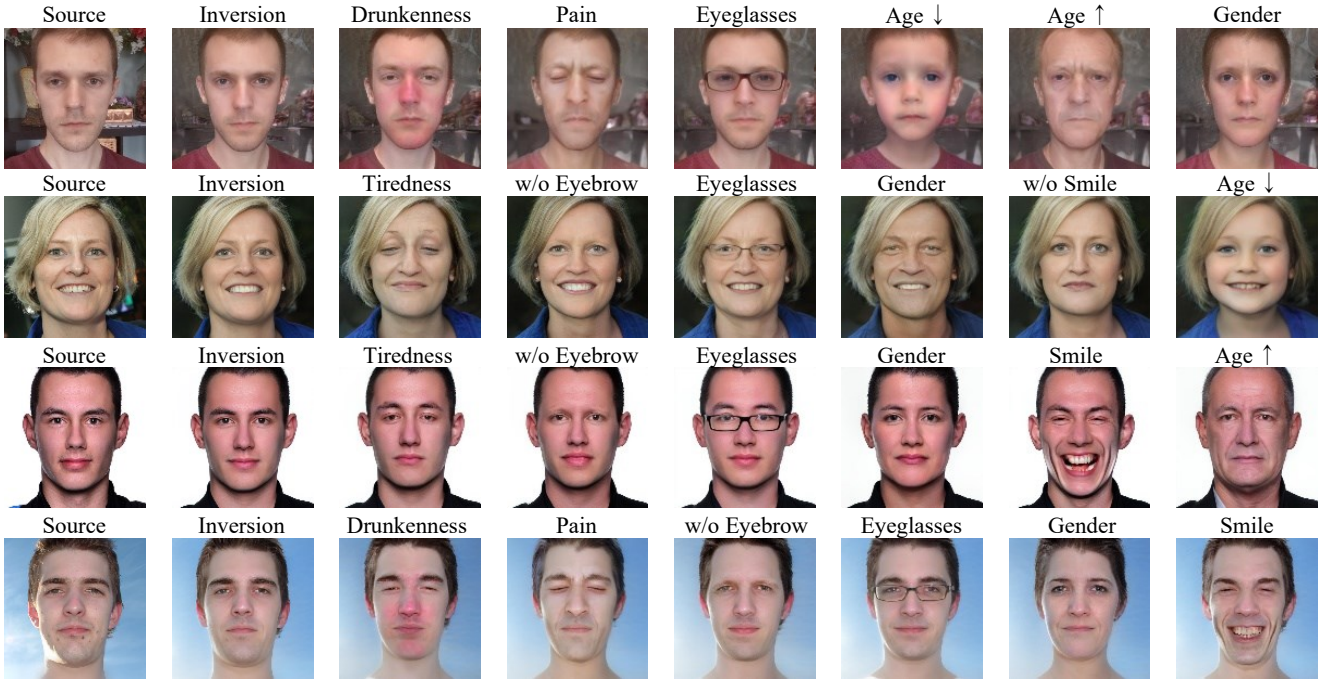

Figure 1. Attribute manipulation on real images. The image in the first column is the source, the image in the second column is the image inverted from the source using pivotal tuning inversion (PTI) [12], and the following six images are the results of each manipulation for an attribute. In the manipulated images, faces are aligned in the frontal direction. The source image in the first row is from the CoNeRF [7] dataset, and the source image in the other rows is from the FFHQ [8] dataset.

C.3. Effect of Few Shots in Quantifier

In this section, we additionally experimented on the number of data samples per group, Ω_a, of the attribute dataset $\mathcal{D}_a$ (See section 3.1). Figure 2 shows face age manipulation experiments using $\Omega_a = 1$ quantifier trained with two different attribute datasets $\mathcal{D}_a$. In both experiments, the age attribute is manipulated as initially intended, but biases appear in each attribute dataset. For example, in the experimental results using attribute dataset A, images manipulated into younger people look more Asian. While in the experimental results using attribute dataset B, images manipulated into older people look more Asian. In addition, there is a problem that the area between the eyes turns red when manipulated to younger. This experimental result implies that the navigator trained using the quantifier shows a biased tendency when the qualifier is trained with a few samples.

We also experimented on the bias according to the number of data samples, Ω_a. Figure 3 shows examples of the age attribute manipulation when varying Ω_a value of attribute dataset $\mathcal{D}_a$. It shows that increasing Ω_a to five or ten significantly reduces the bias compared to when Ω_a is one or two.

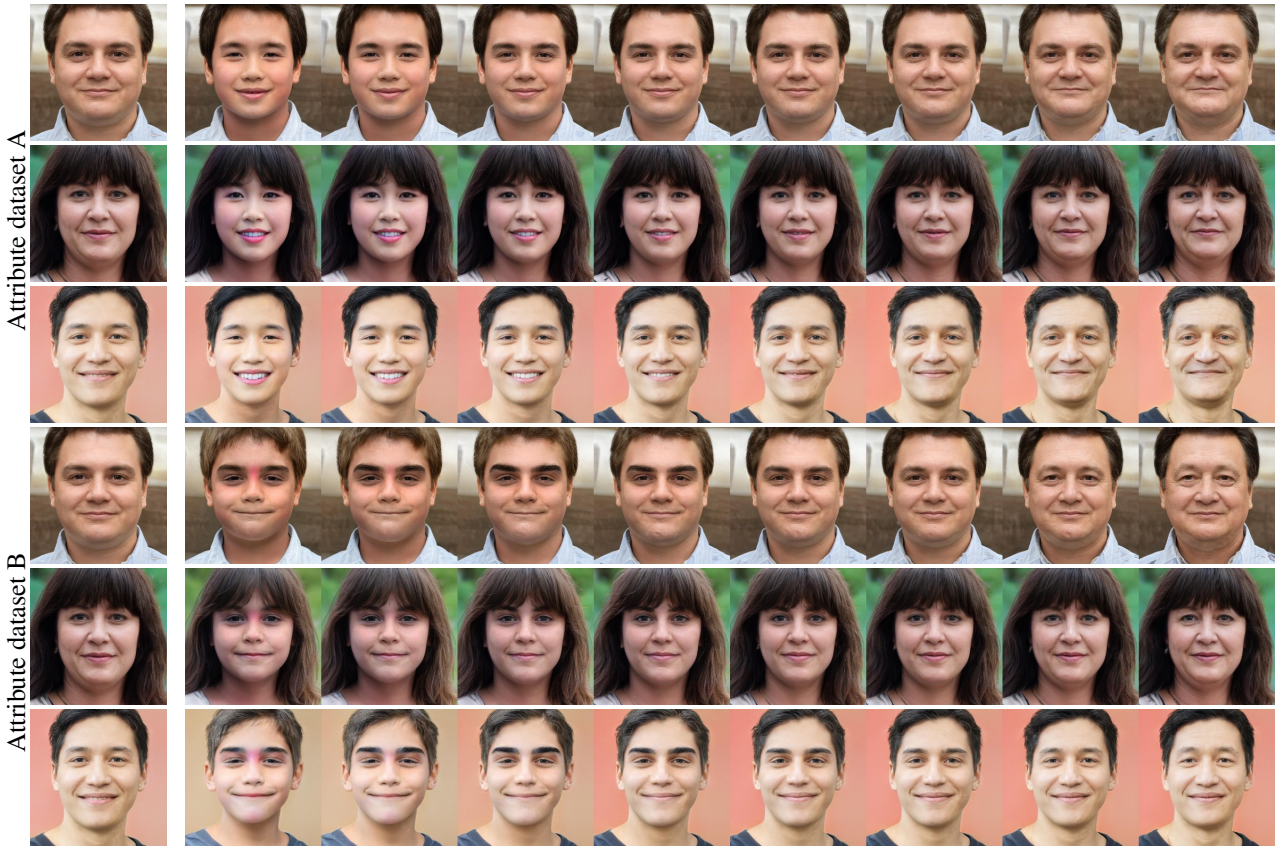

Figure 2. Examples of bias in the face age manipulation using $\Omega_a = 1$ quantifier. For the two different attribute datasets, quantifiers were trained using each attribute dataset, and then navigators were trained using each quantifier. The image in the first column is the source, and the following eight images are the manipulated images with the target quantity increased from 0 to 1.

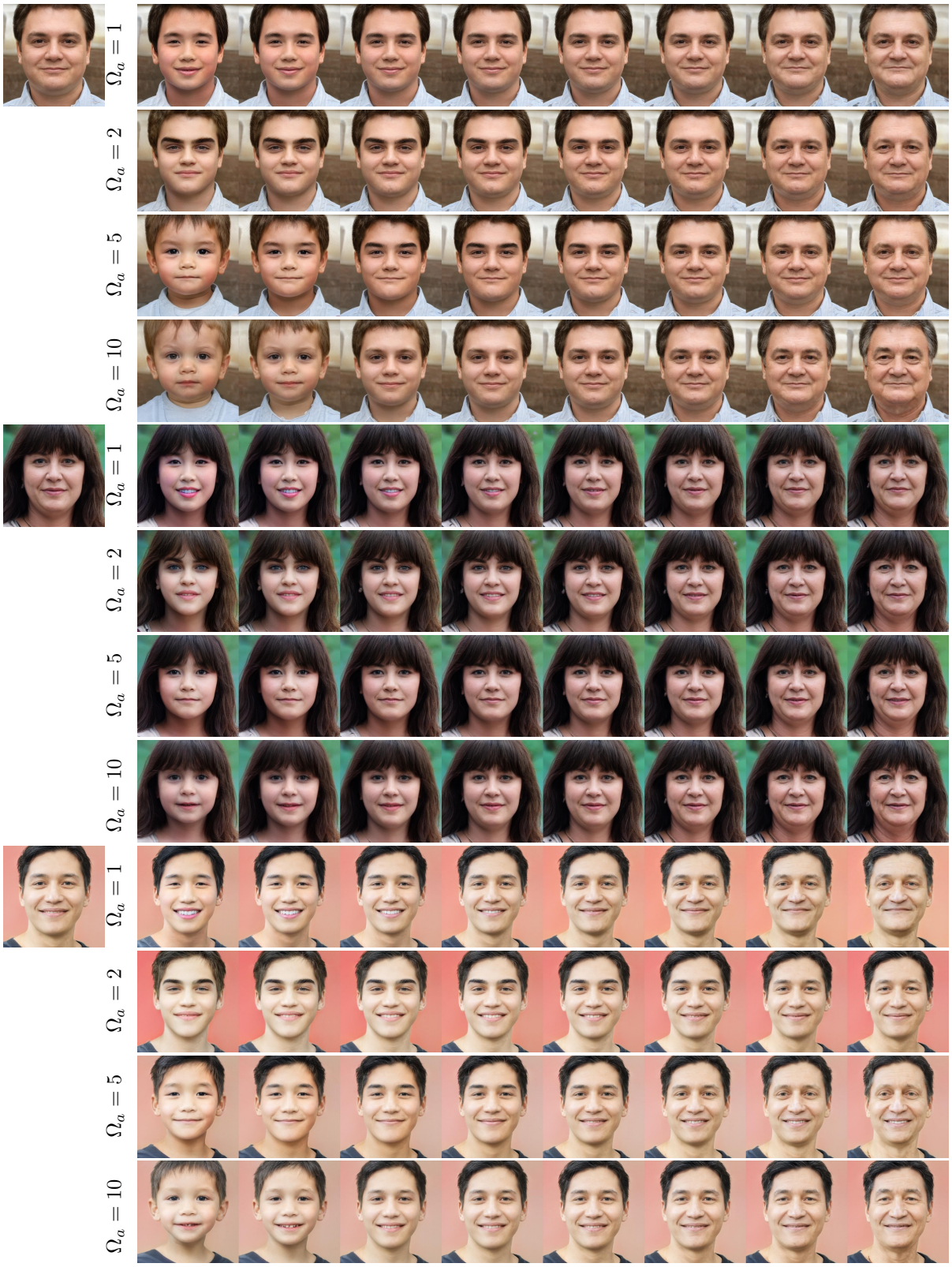

Figure 3. Examples of the face age manipulation depending on the number of images per group, Ω_a, of the quantifier attribute datasets, $\mathcal{D}_a$. $\Omega_a = 1$ examples are the experimental results with attribute dataset A shown in Figure 2.

C.4. Qualitative Comparison in 2D

In this section, we additionally present examples of qualitative comparison in 2D manipulation on StyleGAN2. We compare our method with StyleFlow [2] and Talk-to-Edit [6]. For qualitative comparison, two attributes were tested: age and smile. As shown in Figure 4, our method can manipulate images more independently on the source quantity than the competing methods, which results in better manipulation accuracy. In the case of manipulating images to the youngest age, our method changes both an elderly and a kid into kids (2^{nd} and 9^{th} images in 1^{st} row). However, the competing methods change an elderly into a young adult while changing a kid into a kid (2^{nd} and 9^{th} images in 2^{nd} and 3^{rd} rows). In the case of manipulating a slightly smiling woman to the most unsmiling status, our method changes the woman to expressionless, same as the boy (2^{nd} and 9^{th} images in 4^{th} row). However, StyleFlow shows lower performance than ours; that is, the woman still slightly smiles, unlike the expressionless boy (2^{nd} and 9^{th} images in 4^{th} row). In addition, Talk-to-Edit fails to manipulate the image (2^{nd} image in 5^{th} row).

Figure 4. Qualitative comparison with competing methods for 2d manipulation. FLOW and TALK denote the results of the StyleFlow and Talk-to-Edit methods, respectively. The leftmost column is the source, and the six following images are the manipulated images. Two attributes are experimented with: age (Rows 1–3) and smile (Rows 4–6).

D. Quantitative Manipulation on Various Attributes

We present examples of quantitative manipulation on various attributes in this section. Figure 5, 6, 7, 8, 9, 10, 11, 12, and 13 show examples of the face category's age, smile, gender, eyeglasses, tiredness, drunkenness, pain, eyebrow, and thinness manipulation on EG3D, respectively. Figure 14, 15, and 16 show examples of the cat category's age, eye, and fur manipulation on EG3D, respectively. Figure 17, 18, 19, and 20 show examples of the car category's year, type, chromatic color, and achromatic color manipulation on StyleGAN2, respectively. Figure 21 and 22 show examples of the church category's cloudiness and darkness manipulation on StyleGAN2, respectively. We also invite the reader to watch our supplementary videos, which show fine-grained quantitative manipulation in various cases.

Figure 5. Quantitative manipulation results of the face age attributes. The image in the first column is the source image, and the following eight images are the manipulated images with the target quantity increased from 0 to 1. This layout and the target qualities are the same for the 11 figures in the below (Figure 6 – 16).

Figure 6. Quantitative manipulation results of the face smile attributes.

Figure 7. Quantitative manipulation results of the face gender attributes.

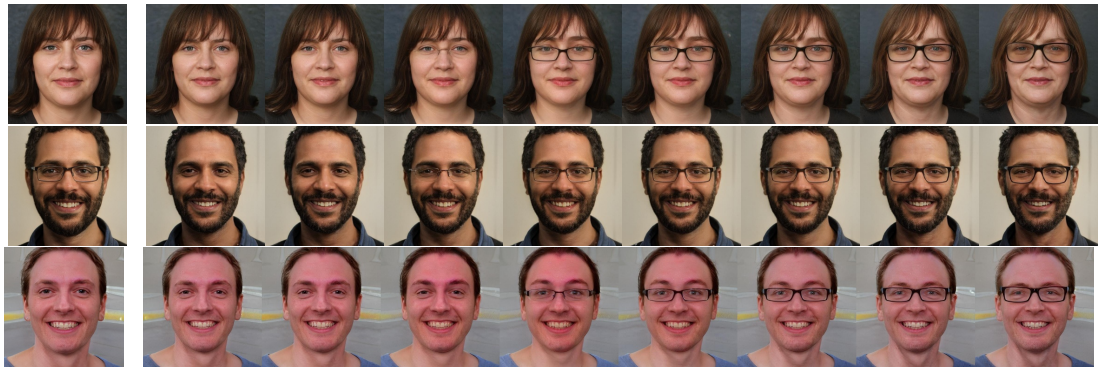

Figure 8. Quantitative manipulation results of the face eyeglasses attributes.

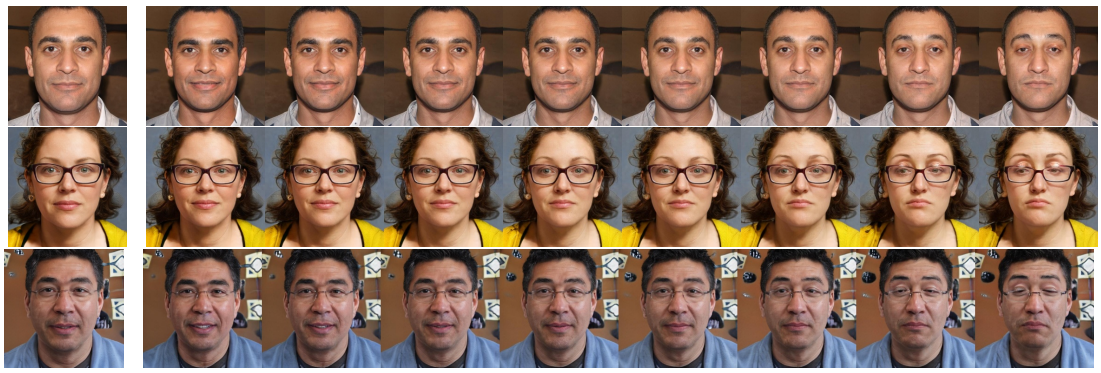

Figure 9. Quantitative manipulation results of the face tiredness attributes.

Figure 10. Quantitative manipulation results of the face drunkenness attributes.

Figure 11. Quantitative manipulation results of the face pain attributes.

Figure 12. Quantitative manipulation results of the face eyebrow attributes.

Figure 13. Quantitative manipulation results of the face thinness attributes.

Figure 14. Quantitative manipulation results of the cat age attributes.

Figure 15. Quantitative manipulation results of the cat eye attributes.

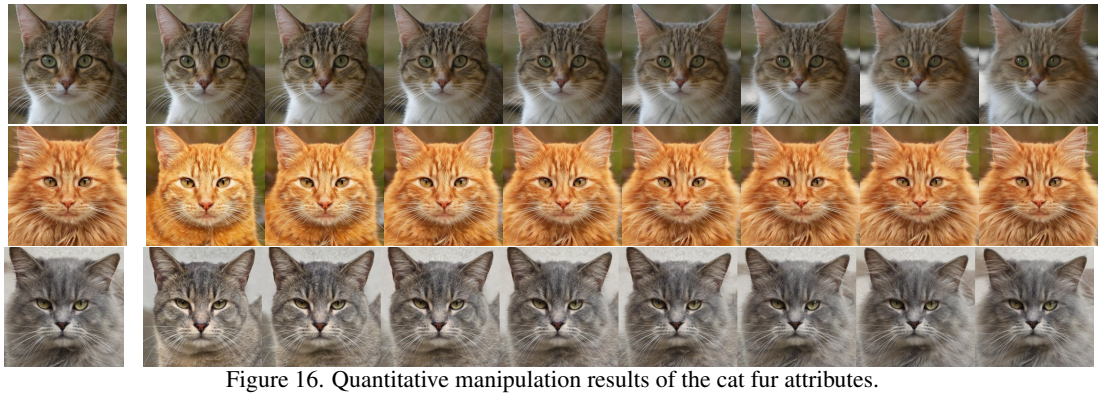

Figure 16. Quantitative manipulation results of the cat fur attributes.

Figure 17. Quantitative manipulation results of the car year attribute. The first row is the samples of the attribute dataset $\mathcal{D}_a$, and the others are the result examples. The image in the first column is the source image, and the following five images are the manipulated images with the target quantity increased from 0 to 1.

Figure 18. Quantitative manipulation results of the car type attribute. The first row is the samples of the attribute dataset $\mathcal{D}_a$, and the others are the result examples. The image in the first column is the source image, and the following five images are the manipulated images with the target quantity increased from 0 to 1.

Figure 19. Quantitative manipulation results of the car chromatic color attribute. The first row is the samples of the attribute dataset $\mathcal{D}_a$, and the others are the result examples. The image in the first column is the source image, and the following four images are the manipulated images with the target quantity increased from 0 to 1.

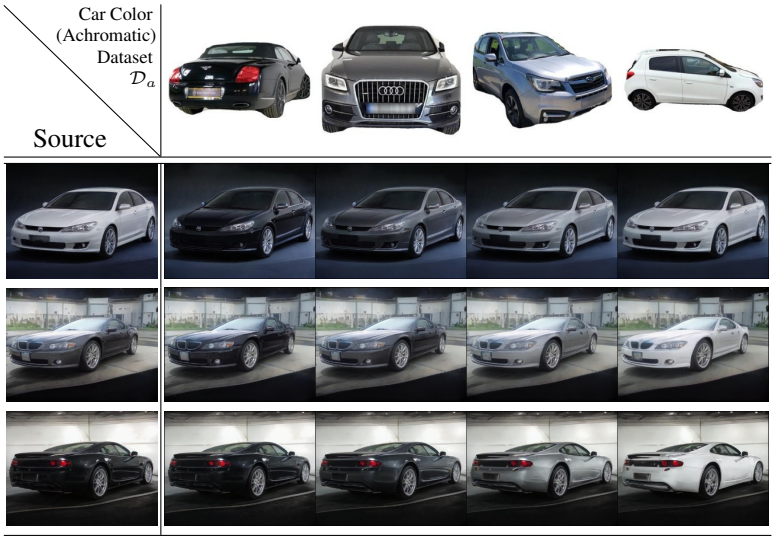

Figure 20. Quantitative manipulation results of the car achromatic color attribute. The first row is the samples of the attribute dataset $\mathcal{D}_a$, and the others are the result examples. The image in the first column is the source image, and the following four images are the manipulated images with the target quantity increased from 0 to 1.

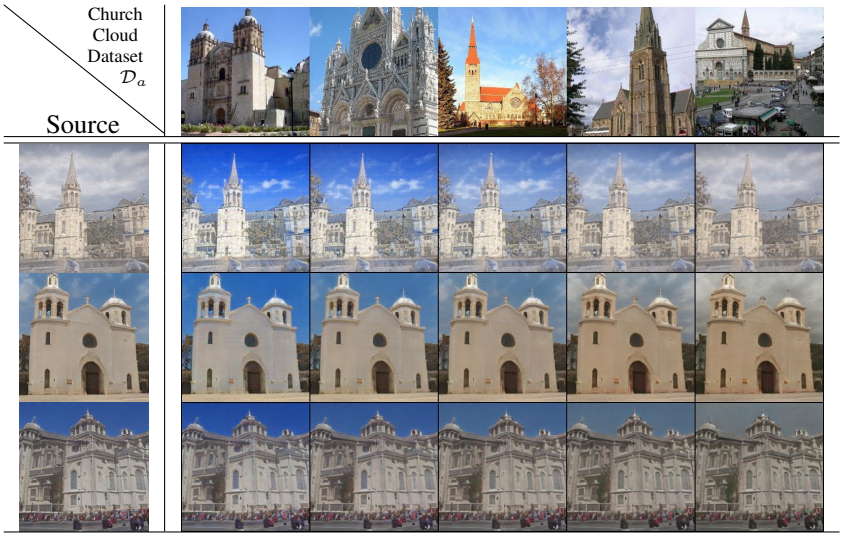

Figure 21. Quantitative manipulation results of the church cloudiness attribute. The first row is the samples of the attribute dataset $\mathcal{D}_a$, and the others are the result examples. The image in the first column is the source image, and the following five images are the manipulated images with the target quantity increased from 0 to 1.

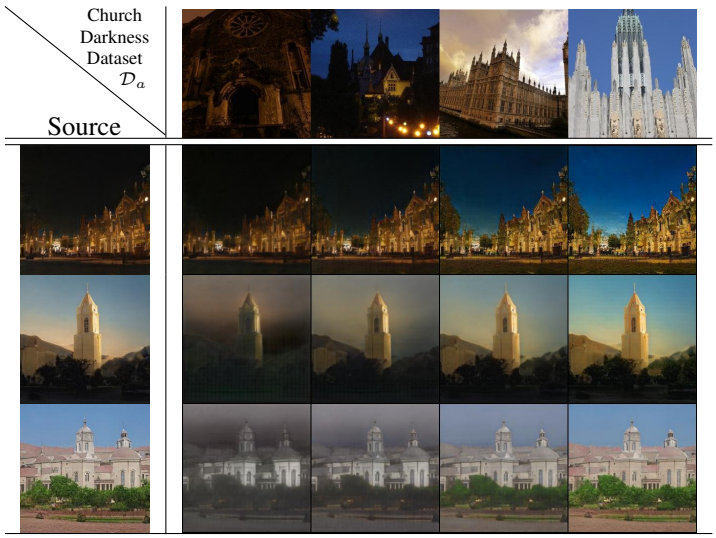

Figure 22. Quantitative manipulation results of the church darkness attribute. The first row is the samples of the attribute dataset $\mathcal{D}_a$, and the others are the result examples. The image in the first column is the source image, and the following four images are the manipulated images with the target quantity increased from 0 to 1.

References

[1] Face++ face detection api. https://www.faceplusplus.com/. Accessed: 2023-02-23. 2

[2] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 6

[3] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *arXiv preprint arXiv:2106.07873*, 2021. 1

[4] Manuel G Calvo, Andrés Fernández-Martín, Guillermo Recio, and Daniel Lundqvist. Human observers and automated assessment of dynamic emotional facial expressions: Kdef-dyn database validation. *Frontiers in psychology*, 9:2052, 2018. 1

[5] Jingming Huang, Bowei Chen, Lan Luo, Shigang Yue, and Iadh Ounis. Dvm-car: A large-scale automotive dataset for visual marketing research and applications. *arXiv preprint arXiv:2109.00881*, 2021. 1

[6] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13799–13808, 2021. 6

[7] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzciński, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18623–18632, 2022. 3

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3

[9] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 1

[10] Microsoft. Azure face. https://azure.microsoft.com/en-in/services/cognitive-services/face/., 2020. 2

[11] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *European Conference on Computer Vision*, pages 739–755. Springer, 2020. 1

[12] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 3

[13] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 1

[14] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019. 1