L2M-GAN: Learning to Manipulate Latent Space Semantics for Facial Attribute Editing – Supplementary Material –

Guoxing Yang[1,2] Nanyi Fei[1] Mingyu Ding[3] Guangzhen Liu[1] Zhiwu Lu[1,2] Tao Xiang[4]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods
[3]The University of Hong Kong [4]University of Surrey, UK
luzhiwu@ruc.edu.cn

This document is organized as follows. In Sec. 1, we present the details of network architecture of our L2M-GAN. In Sec. 2, we provide the visual results on AFHQ [2] to show the generalization ability of our L2M-GAN. In Sec. 3, we present the comparison with StarGAN V2 [2]. In Sec. 4, we give more quantitative and qualitative results for facial attribute editing: (1) comparative results on gender and eyeglasses attributes by comparing our L2M-GAN with the state-of-the-art methods; (2) visual results on editing five attributes by our L2M-GAN.

Layer	Resample	Norm	Output Shape
Image x	–	–	$256 \times 256 \times 3$
Conv3×3	–	–	$256 \times 256 \times 64$
ResBlock	AvgPool	IN	$128 \times 128 \times 128$
ResBlock	AvgPool	IN	$64 \times 64 \times 256$
ResBlock	AvgPool	IN	$32 \times 32 \times 512$
ResBlock	AvgPool	IN	$16 \times 16 \times 512$
ResBlock	AvgPool	IN	$8 \times 8 \times 512$
ResBlock	–	IN	$8 \times 8 \times 512$
ResBlock	–	IN	$8 \times 8 \times 512$
ResBlock	–	AdaIN	$8 \times 8 \times 512$
ResBlock	–	AdaIN	$8 \times 8 \times 512$
ResBlock	Upsample	AdaIN	$16 \times 16 \times 512$
ResBlock	Upsample	AdaIN	$32 \times 32 \times 512$
ResBlock	Upsample	AdaIN	$64 \times 64 \times 256$
ResBlock	Upsample	AdaIN	$128 \times 128 \times 128$
ResBlock	Upsample	AdaIN	$256 \times 256 \times 64$
LReLU	–	–	$256 \times 256 \times 64$
Conv1×1	–	–	$256 \times 256 \times 3$

Table 1. Network architecture of generator.

1. Details of Network Architecture

In this section, we introduce the detailed network architecture of our L2M-GAN, which consists of generator, style transformer, style encoder, and discriminator.

Layer	Type	Activation	Output Shape
Style Code s	Shared	–	64
Linear	Shared	ReLU	64
Linear	Shared	ReLU	64
Linear	Shared	–	64

Table 2. Network architecture of decomposer.

Layer	Type	Activation	Output Shape
Style Code s_{re}	Shared	–	64
Linear	Unshared	ReLU	64
Linear	Unshared	ReLU	64
Linear	Unshared	–	64

Table 3. Network architecture of domain transformer.

Generator. As shonw in Table 1, our generator consists of five downsampling blocks with instance normalization (IN), four immediate blocks, and five upsampling blocks for input images at the resolution of 256×256. All blocks are residual blocks as in StarGAN v2[2]. We also use the adaptive instance normalization (AdaIN) [4, 5] for upsampling blocks, where a style code provides scaling and shifting vectors through learned affine transformations.

Style Transformer. Our style transformer consists of decomposer and domain transformer. The network architectures of decomposer and domain transformer are shown in Table 2 and Table 3, respectively. Concretely, decomposer is a multi-layer perceptron (MLP) containing three fully connected layers, which are shared among all domains. Moreover, domain transformer consists of $\mathcal{K}$ unshared MLPs, where $\mathcal{K}$ denotes the number of domains. Each MLP in domain transformer contains three specific fully connected layers for each domain. In our experiments, the dimension of the style code is set to 64.

Style Encoder. As shown in Table 4, our style encoder consists of a CNN with $\mathcal{K}$ output branches, where $\mathcal{K}$ is the num-

Layer	Resample	Norm	Output Shape
Image x	–	–	$256 \times 256 \times 3$
Conv3×3	–	–	$256 \times 256 \times 64$
ResBlock	AvgPool	IN	$128 \times 128 \times 128$
ResBlock	AvgPool	IN	$64 \times 64 \times 256$
ResBlock	AvgPool	IN	$32 \times 32 \times 512$
ResBlock	AvgPool	IN	$16 \times 16 \times 512$
ResBlock	AvgPool	IN	$8 \times 8 \times 512$
ResBlock	–	IN	$4 \times 4 \times 512$
LReLU	–	–	$4 \times 4 \times 512$
Conv4×4	–	–	$1 \times 1 \times 512$
LReLU	–	–	$1 \times 1 \times 512$
Reshape	–	–	512
Linear * $\mathcal{K}$	–	–	$\mathcal{D} * \mathcal{K}$

Table 4. Network architecture of style encoder and discriminator. Note that $\mathcal{D}$ denotes the output dimension and $\mathcal{K}$ denotes the number of domains.

ber of domains. Six residual blocks with instance normalization (IN) are shared among all domains and one specific fully connected layer is used for each domain. We adopt a convolution layer after the last residual blocks to get the feature vector instead of average polling. The output dimension $\mathcal{D}$ is the dimension of style code (i.e., $\mathcal{D}$=64).

Discriminator. Our discriminator is a multi-task discriminator, which has the same architecture as style encoder except the output dimension. The net architecture of discriminator is also shown in Table 4 and the output dimension $\mathcal{D}$ is set to 1 for binary classification. For each domain, we use a fully connected layer for real/fake classification.

2. Visual Results on AFHQ

AFHQ [2] is a high-quality dataset of animal faces, consisting of 15,000 images at 512 × 512 resolution. We separate AFHQ into three domains of cat, dog, and wild as in StarGAN V2 [2], and resize all the images to 256 × 256. The visual results obtained by our L2M-GAN are shown in Figure 1. We can see that our L2M-GAN also transfers the input images into the target domain correctly and preserves the other information well even on this non-face dataset, which further demonstrates the effectiveness and robustness of our L2M-GAN.

3. Comparison with StarGAN V2

We did consider StarGAN V2 [2] as a baseline but found that it is clearly inferior to the proposed model as well as StarGAN V1 for facial attribute editing. In particular, a good facial attribute editing model should meet two requirements: attribute correctness and irrelevance preservation. But StarGAN V2 tends to make too many unintended changes to the original images (see identity and hair changes in Figure 2), thus not meeting the second requirement. Therefore, we do not consider StarGAN V2 as one the comparative methods in the main paper.

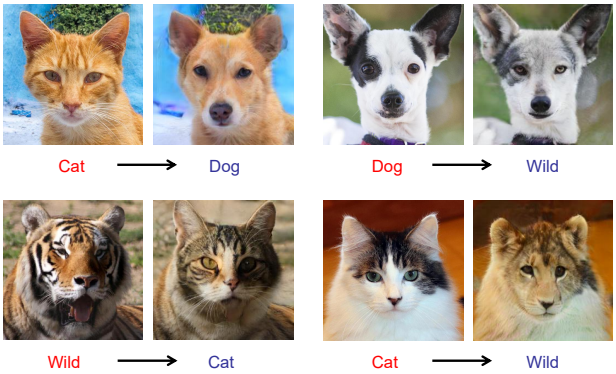

Figure 1. Visual results on AFHQ by our L2M-GAN.

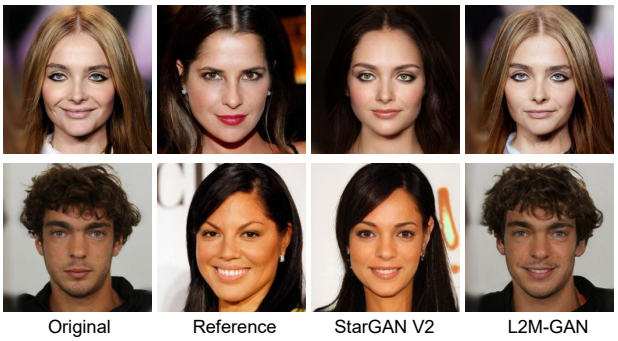

Figure 2. Visual results of editing smiling by StarGAN V2.

4. Additional Results

In this section, we provide more quantitative and qualitative results for facial attribute editing: (1) comparative results on gender and eyeglasses attributes by comparing our L2M-GAN with the state-of-the-art methods; (2) visual results on editing five attributes by our L2M-GAN.

Comparative Results on Gender and Eyeglasses Attributes. We compare our L2M-GAN with the state-of-the-art methods on two additional attributes: **Gender** and **Eyeglasses**. We adopt attribute manipulation accuracy and quality of generated images for quantitative evaluation. The quantitative results on the two attributes are shown in Table 5 and Table 6, respectively. We have the following observations: (1) Our L2M-GAN outperforms the other methods on attribute manipulation accuracy on both attributes. (2) PA-GAN achieves high accuracy on a local attribute (eg. **Eyeglass**) but still suffers from insufficient modification on a global attribute (eg. **Gender**), which results in relative low attribute manipulation accuracy. (3) StarGAN achieves relative high accuracy on both attribute at the cost of image quality degradation. (4) Our L2M-GAN obtains the best FID for adding and removing eyeglasses and transferring

Method	FID (+)	FID (-)	FID (avg)	Acc (att)
StarGAN [1]	74.0	102.6	88.3	84.8%
CycleGAN [8]	41.9	41.8	41.9	90.9%
ELEGANT [7]	83.3	86.7	85.0	81.6%
PA-GAN [3]	77.1	97.4	87.3	79.6%
InterFaceGAN* [6]	59.7	66.9	63.3	81.3%
L2M-GAN (ours)	**34.1**	**37.6**	**35.9**	**94.9%**

Table 5. Quantitative results for facial attribute editing on the specific attribute: **Gender**. FID (+) (or FID (-)) denotes the FID score for transferring female to male (or male to female), and FID (avg) denotes the simple average of FID (+) and FID (-). Acc (att) denotes the attribute manipulation accuracy.

Method	FID (+)	FID (-)	FID (avg)	Acc (att)
StarGAN [1]	87.8	100.7	94.3	93.7%
CycleGAN [8]	37.4	81.4	59.4	98.0%
ELEGANT [7]	60.8	105.4	83.1	89.5%
PA-GAN [3]	74.2	110.0	92.1	91.6%
InterFaceGAN* [6]	70.7	89.2	80.0	98.4%
L2M-GAN (ours)	**31.5**	**76.6**	**54.0**	**98.5%**

Table 6. Quantitative results for facial attribute editing on the specific attribute: **Eyeglasses**. FID (+) (or FID (-)) denotes the FID score for adding (or removing) eyeglasses, and FID (avg) denotes the simple average of FID (+) and FID (-). Acc (att) denotes the attribute manipulation accuracy.

gender to opposite one, which indicates that it can generate images with highest quality on both attributes.

The qualitative results for editing the gender and eyeglasses attributes are shown in Figure 3 and Figure 4, respectively. From these two figures, we can observe that: (1) StarGAN and CycleGAN can generate images with correct attribute but still tend to generate blurs and artifacts. They even change the style of images in some cases. (2) Elegant can not transfer gender from reference images correctly, which results in much blurs and artifacts. It can transfer eyeglasses from reference but still generate much blurs and artifacts as a result of entanglement in the latent space. (3) PA-GAN tends to preserve the irrelevant regions well because of region attention but also suffers from insufficient modification on both attributes. (4) InterfaceGAN* generates high-quality images but always changes the identity information of the input image due to not not considering identity during factorization. (5) Our L2M-GAN makes correct attribute manipulation both local attribute and global attribute and produces high-quality images. Importantly, the other attributes and identity information of the input image are preserved well in the generated images, which demonstrates that our L2M-GAN can change the attribute-relevant information correctly whilst preserving the attribute-irrelevant information well.

Qualitative Results on Editing Five Attributes. In addition to Figure 1 of the main paper, we provide additional qualitative results on editing five attributes obtained by our L2M-GAN in Figure 5. We can see that our L2M-GAN generally transfers the input images into target domain correctly with high-quality results, which makes desired attributes appear on the generated images correctly and preserves irrelevant information well in the mean time.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (61976220 and 61832017), and Beijing Outstanding Young Scientist Program (BJJWZYJH012019100020098). Zhiwu Lu is the corresponding author.

References

[1] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018. 3, 4, 5

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8185–8194, 2020. 1, 2

[3] Zhenliang He, Meina Kan, Jichao Zhang, and Shiguang Shan. PA-GAN: Progressive attention generative adversarial network for facial attribute editing. *arXiv preprint arXiv:2007.05892*, 2020. 3, 4, 5

[4] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, 2017. 1

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1

[6] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *CVPR*, pages 9240–9249, 2020. 3, 4, 5

[7] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. ELEGANT: Exchanging latent encodings with GAN for transferring multiple face attributes. In *ECCV*, pages 172–187, 2018. 3, 4, 5

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 3, 4, 5

Figure 3. Qualitative results for facial attribute editing on the specific attribute: **Gender**. The first column shows the real source/input images. The other columns from left to right are the editing results of StarGAN [1], CycleGAN [8], ELEGANT [7], PA-GAN [3], InterfaceGAN* [6], and our L2M-GAN. Better viewed on-line in color and zoomed in for details.

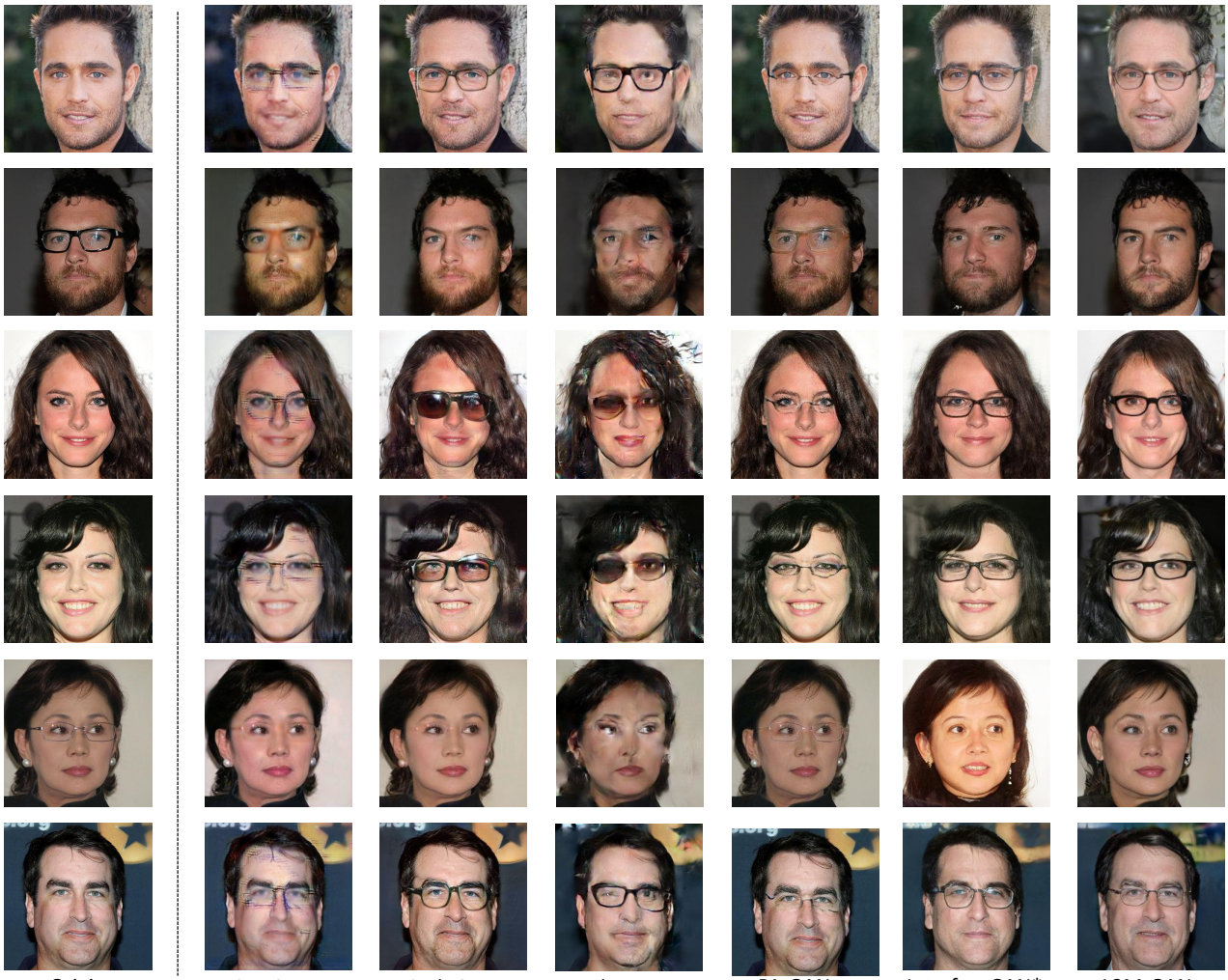

Figure 4. Qualitative results for facial attribute editing on the specific attribute: **Eyeglasses**. The first column shows the real source/input images. The other columns from left to right are the editing results of StarGAN [1], CycleGAN [8], ELEGANT [7], PA-GAN [3], InterfaceGAN* [6], and our L2M-GAN. Better viewed on-line in color and zoomed in for details.

Figure 5. Attribute editing results by our L2M-GAN on CelebA-HQ. The first column shows the real source images, and each of the other columns shows the results of editing a specific attribute. Each edited image has an attribute value opposite to that of the source one.