Computational Visual Media
DOI 10.1007/s41095-xxx-xxxx-x Vol. x, No. x, month year, xx–xx

Research Article

Mask-aware Photorealistic Face Attribute Manipulation

Ruoqi Sun[1], Chen Huang[2], Hengliang Zhu[1], and Lizhuang Ma[1](✉)

Abstract The technique of face attribute manipulation has found increasing applications, but remains challenging with the restriction of editing the attributes while preserving its unique details. In this paper, we introduce our method named the Mask-Adversarial AutoEncoder (M-AAE) which combines the Variational AutoEncoder (VAE) and Generative Adversarial Network (GAN) for photorealistic image generation. We propose the partial dilated layers to modify a modest amount of pixels in the feature maps of an encoder, changing the attribute strength continuously without hindering global information. Our training objectives of VAE and GAN are reinforced by the supervision of face recognition loss and cycle consistency loss for faithful preservation of face details. Moreover, we generate facial masks to enforce background consistency, which allows our training to focus on foreground face rather than background. Experimental results demonstrate that our method, can generate high-quality images with varying attributes and outperform the existing methods in detail preservation.

Keywords Face Attribute Manipulation; Generative Adversarial Network(GAN); Variational AutoEncoder(VAE); Partial Dilated Layers; Photorealistic Mechanism.

1 Introduction

The task of face attribute manipulation is to edit the face attributes shown in an image, e.g., hair color, facial expression, age and so on. It has a wide range of applications, such as data augmentation and age-invariant face verification [3, 26, 29, 38]. Essentially, this is an image generation problem. With the advent of generative adversarial networks (GANs), the quality of generated images improves over time [7, 41]. The family of GAN methods can be mainly divided into two categories: one with noise input [24, 37] and another conditioned on input images [2, 4, 23]. Our method falls into the second category, aiming to change the face attributes in the input image while preserving high-frequency details.

Normally, the neural network generate the result images by manipulating all the pixels of the input image. However, unlike the style translation task [5, 19], the attribute manipulation one is more challenging due to the restriction of only modifying some image features while keeping others unchanged (including the image background). In this paper, we improve the quality of the manipulated images from three aspects: the concentration of attribute manipulation, the preservation of facial details and the photorealistic mechanism.

The concentration of attribute manipulation. The manipulation method aims to focus on modifying the target attributes while keep the common feature unchanged. One simple choice to achieve this goal is to use the conditional GAN framework [24, 39], which concatenates the input image with an one-hot attribute vector to encode the desired manipulation. Another option is to directly learn the image-to-image translation with respect to attributes. CycleGAN [42] learns such translation rule from unpaired images with a cycle consistency constraint. However, such global transformation can neither guarantee common feature preservation, nor make a continuous change in the

[1] Shanghai Jiao Tong University, Shanghai, 200240, China. E-mail: Ruoqi Sun, ruoqisun7@sjtu.edu.cn; Hengliang Zhu, hengliang_zhu@sjtu.edu.cn; Lizhuang Ma, malz@cs.sjtu.edu.cn(✉).

[2] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA. E-mail: chenhuang@apple.com.

Manuscript received: 2020-11-03; accepted: 2020-12-21.

attribute strength.

The preservation of facial details. Although achieving promising results, the above methods have one common drawback — there exists no mechanisms to keep the unique facial traits while editing the whole images. It may still change the non-targeted features beyond the background, which is not preferred. We especially note the importance of keeping the background unchanged since it is often observed to be changed along with the foreground face. This suggests some efforts of face attribute manipulation are wasted on the irrelevant regions. Moreover, the post-process of overriding generated background with the original one by a background mask would be less preferred, as it needs better handling along the boundaries to avoid visible seams.

The photorealistic mechanism. The realistic of the generated image is one of the most important measurement of the image generation algorithm, including the fidelity of face features, the clarity of images and so on. Since the features are various, different methods are proposed to fit the special tasks. The method of [39] provides a partial remedy by feeding the face images before and after attribute manipulation into a face recognition network and penalizing their feature distance. This provides a good way to preserve facial identify information. Recently, UNIT method [20] uses generative adversarial networks (GANs) and variational autoencoders (VAEs) for robust modeling of different image domains. Then the cycle consistency constraint is also applied to learn domain translation effectively. The method of [32] proposes to only learn the residual image before and after attribute manipulation by using two transformation networks, one for attribute manipulation and another for its dual operation. However, the methods mentioned above focus on the single task.

In this paper, we train a neural network to simultaneously manipulate the target attributes of a face image and keep its background untouched. Firstly, we propose the patial dilated layer to modify the minimum number of feature map pixels from our encoder. It allows us to maximally preserve the global image information and enables attribute change in a continuous manner. Secondly, we feed the background mask into the network to coherently penalize their differences before and after face attribute manipulation. Finally, our method is based on the VAE-GAN framework [20, 39] for strong modeling of photorealistic images. To avoid loss of unique facial details during attribute editing, we employ a face recognition loss and a cycle consistency loss (to ensure image consistency after two inverse manipulations). The proposed method is named as Mask-Adversarial AutoEncoder (M-AAE) and the experimental results demonstrate its effectiveness.

In summary, the contributions of this paper are as follows. (1) We present the partial dilated layers to modify a modest amount of pixels in our learned feature maps to realize continuous manipulation of face attributes. (2) We propose a Mask-Adversarial AutoEncoder (M-AAE) strategy to ensure faithful facial detail preservation as well as background consistency. (3) We combine the GAN, VAE, mask loss, id loss, cycle consistency loss to generate the photorealistic facial images. The proposed method achieves state-of-the-art performance in photorealistic attribute manipulation.

2 Related work

Face attribute manipulation Considerable progress has been made on face attribute manipulation [1, 8, 9, 12, 17, 18]. Most methods of face attribute manipulation are based on generative models. There are two main groups of these methods: the one with extra input vector [4, 6, 28, 39], and the other group that directly learn the image-to-image translation along attributes [20, 42]. The first group often takes an attribute vector as the guidance for manipulating the desired attribute. The CAAE method [39] concatenates the one-hot age label with latent image features to be fed into the generator for age progression purposes. StarGAN [4] takes the one-hot vector to represent domain information for ”domain transfer”. However, such global transformation based on external code usually cannot well preserve the facial details after attribute manipulation. The second group of methods only operate in image domains and learn the image-to-image translation directly. The CycleGAN [42] and UNIT method [20] are such examples, supervised by a cycle consistency loss that requires the manipulated image can be mapped back to the original image. [32] further proposed to only learn the residual image before and after attribute manipulation, which can be easier and lead to higher-quality image prediction. Unfortunately, these methods still have difficulty of manipulating the target attribute while keeping others unchanged.

Image generation algorithm The Variational

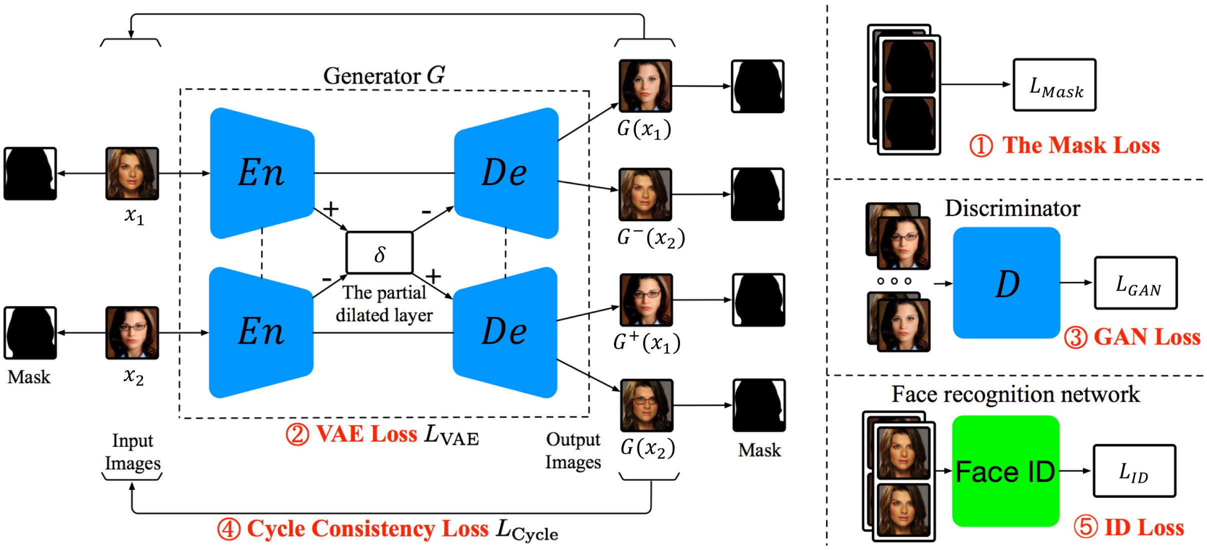

Fig. 1 Framework of the proposed Mask-Adversarial AutoEncoder (M-AAE) method. The encoder-decoder $De(En(x))$ of VAE for input image x is treated as the generator $G(x)$ of GAN, with a discriminator $D(\cdot)$ tells fake from real. We manipulate attributes by modifying the encoded features $En(x)$ by a relative value $\pm\delta$, and train using image pairs with opposite face attributes. Moreover, the encoded features $En(x)$ come from the partial dilated layer. Our training is supervised by 5 loss functions to both preserve facial details and ensure background consistency (see text for details). We test only using the generator $G(\cdot)$.

AutoEncoder (VAE) [16] and Generative adversarial network (GAN) [7] are the backbone for image generation tasks nowadays, such as image reconstruction [10, 31, 34, 36], image synthesis [11, 30, 37] and image translation [14, 25, 33]. In VAE, the encoder maps images into a latent feature space which is then mapped back to the image domain through a decoder. The latent space contains the global features extracted for input images. The more recent GAN consists of the generator and the discriminator networks to play a min-max game. Specifically, the generator tries to produce synthesized images to fool the discriminator that distinguishes the synthesized images from real ones. GAN-based methods have shown remarkable results in image generation, and many improvements followed up. DCGAN [30] trains stable in a purely convolutional setting, while CGAN [24] generates visually compelling images conditioned on extra input like class labels. CycleGAN [42] and UNIT method [20] introduce a cycle consistency loss to learn between any image domains with even unpaired images. There is a recent trend to combine the GAN with a VAE for robust image modeling. For example, [18] combined GAN and VAE by collapsing the VAE decoder and GAN generator into one. One can tweak the generated images by manipulating features in the latent feature space. Such joint VAE-GAN model is also applied in the works of [20, 39] for image translation. Recently, it is possible to generate the human face images in high quality by using GANs [35, 40]. Kim [13] proposed to utilize the GAN to transfer the full 3D head expression from the source actor to the target actor in the video. This paper uses the VAE-GAN model for face attribute manipulation, and proposes a working method to modify latent VAE features so as to change facial attributes but not irrelevant details.

3 Methodology

Our goal is to manipulate the attribute of an input face image and generate a new one, e.g., to change the hair color from black to yellow. However, it is difficult to generate photorealistic images as well as keep the face faithful, i.e., the generated image should look real and have its unique details preserved including the background. To address these challenges we propose a Mask-Adversarial AutoEncoder (M-AAE) method, as will be detailed as follows.

3.1 Framework Overview

Our M-AAE method is based on the VAE-GAN framework, as shown in Fig. 1. The encoder-decoder $De(En(x))$ of VAE for input image x is treated as GAN's generator $G(x)$. The discriminator $D(\cdot)$ of GAN tells the generated image $G(x)$ apart from real images. To manipulate attributes of input image x, we design a simple but effective mechanism to uniformly modify the

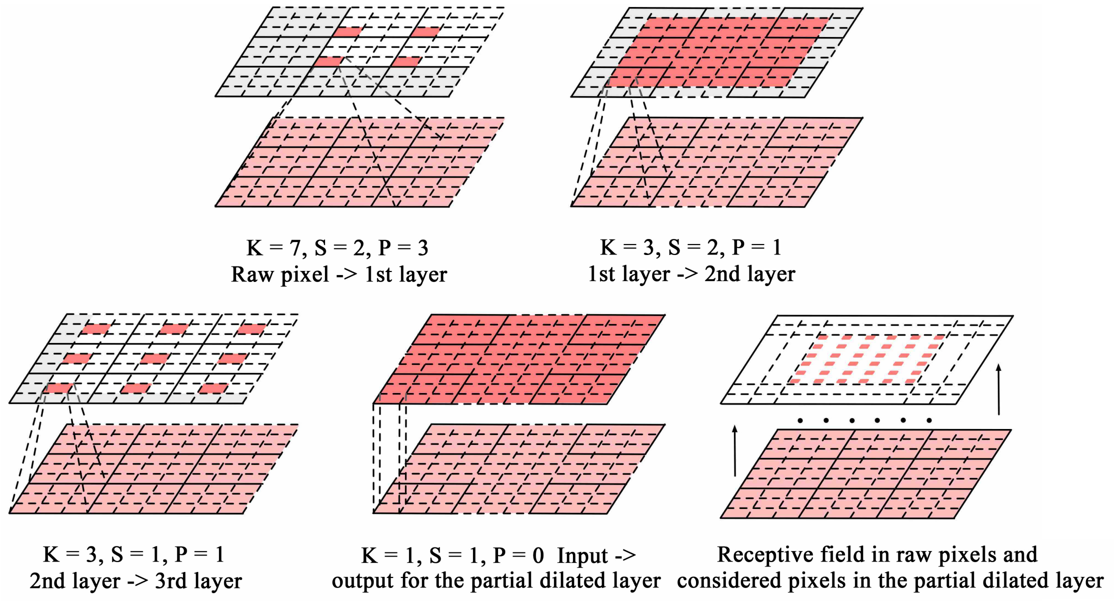
K = 7, S = 2, P = 3
Raw pixel -> 1st layer
K = 3, S = 2, P = 1
1st layer -> 2nd layer
K = 3, S = 1, P = 1
2nd layer -> 3rd layer
K = 1, S = 1, P = 0 Input ->
output for the partial dilated layer
Receptive field in raw pixels and
considered pixels in the partial dilated layer

Fig. 2 The receptive fields of the four kinds of partial dilated layers (from bottom to top) of our Encoder (K, S, P denotes the kernel size, stride and padding, respectively, see Table 1 for details). The padding area in each layer during recursive calculation is not counted as the receptive field. The rightmost subfigure in the second row shows the global receptive field in raw pixels and modified pixels in the partial dilated layer. It demonstrates our goal to find the minimum number of feature map pixels at the partial dilated layer, whose receptive field covers the whole image in image domain.

encoded features $En(x)$ by a relative value $\pm\delta$, which is fed into the decoder to control the attribute strength present in output $G^{+}(x)/G^{-}(x)$.

We propose the partial dilated layers to manipulate the face feature continuously while preserving the consistent of the global features during the manipulation process. Furthermore, the mask-aware method is utilized to separate the foreground and the background of the input images. Thus, the method can focus on the foreground images and manipulate the chosen feature only in the foreground, which can reduce the influence of the background. Modifying the image features by using a small number of the features instead of modifying the whole pixels of the images can protect the image features. The proposed losses focus on different aspects, including the identification and the age of the faces, the clarity of the images and so on. The combination of these losses achieves better performance on improving the quality of the image.

3.2 The Partial Dilated Layers for Attribute Manipulation

To manipulate face attributes, rather than take a one hot attribute vector as in [4, 39], we choose to modify the hidden features in our encoder to be able to continuously change the attribute strength. One intuitive way is to uniformly increase or decrease the responses of the entire feature map by a relative value δ. We empirically observed a global change of image tone by doing this. Instead, we propose to only modify a minimum number of latent feature map pixels in the CNN whose receptive field covers the whole image in image domain. Fig. 2 illustrates how to find such minimum pixels at the partial dilated layer(the last layer of the encoder) recursively from bottom layer. In this way, the image-level manipulation can be operated efficiently with modest feature modification. More importantly, we will avoid a huge loss of image information. Our experiments will show our efficacy in information preservation during attribute manipulation.

In practice, the relative value δ is chosen as half the value range of the feature map pixels for reversing one particular attribute ($\delta \approx 5$ in our scenario). Then such modified features are fed into the decoder to generate output image $G^{+}(x)$ or $G^{-}(x)$ with strengthened or weakened attribute. For instance, adding the δ means strengthen the face attribute, otherwise weaken it. We change the value in the training process(when we apply the cycle consistency loss) to force saving the strength information in it.

Tab. 1 The network architecture of our Encoder, Decoder and GAN discriminator (channel number, kernel size).

Encoder $En(\cdot)$	Decoder $De(\cdot)$	GAN discriminator $D(\cdot)$
Conv2d (64,7×7) + LeakyReLU	Residual Block (512,1×1)	Conv2d (64,3×3) + LeakyReLU
Conv2d (128,3×3) + LeakyReLU	Residual Block (512,1×1)	Conv2d (128,3×3) + LeakyReLU
Conv2d (256,3×3) + LeakyReLU	Residual Block (512,1×1)	Conv2d (256,3×3) + LeakyReLU
Residual Block (512,1×1)	Conv2d (256,3×3) + LeakyReLU	Conv2d (512,3×3) + LeakyReLU
Residual Block (512,1×1)	Conv2d (128,3×3) + LeakyReLU	Conv2d (1024,3×3) + LeakyReLU
Residual Block (512,1×1)	Conv2d (64,7×7) + LeakyReLU	Conv2d (1,2×2) + Sigmoid

3.3 The Mask-aware Algorithm for Facial Detail Preservation

In some cases, we observed the image background would change along with the foreground face by previous attribute manipulation methods. This is not visually pleasing and also suggests some manipulation efforts are wasted in wrong regions. We claim that pasting the original background around the manipulated face is not ideal. Because the pixels in the final image coming from images in different distribution seems incompatible. More importantly, it is better to mask out background at the algorithm level to focus our manipulation efforts on foreground face. As a side effect, the background gets unchanged as well. Here we propose the **mask loss** to learn to change the foreground face attribute and keep background the same in a coherent way. We generate a facial mask (thus background mask as well) by using FCN [22], and penalize the background difference between input x and generated $G(x)$:

$$\mathcal{L}_{\text{Mask}} = ||\text{Mask}(G(x)) - \text{Mask}(x)||_1, \tag{1}$$

where $\text{Mask}(\cdot)$ is the mask-out operator using the generated background mask. Note the background mask of input x is shared for both input x and output $G(x)$. We do not generate a separate mask for $G(x)$ which leads to inconsistent penalty.

3.4 The Photorealistic Mechanism

VAE loss The VAE consists of an encoder that maps an image x to a latent feature $z \sim En(x) = q(z|x)$ and a decoder that maps z back to image space $x' \sim De(z) = p(x|z)$. The VAE regularizes the encoder by imposing a prior over the latent distribution $p(z)$, where $z \sim \mathcal{N}(0, I)$ is often assumed to have a Gaussian distribution. VAE also penalizes the reconstruction error between x and x', and has loss function:

$$\begin{aligned} \mathcal{L}_{\text{VAE}} &= \lambda_1 \text{KL}(q(z|x)||p(z)) \\ &-\lambda_2 E_{x \sim p_{\text{data}}(x)}[\log\ p(x'|x)], \end{aligned} \tag{2}$$

where λ_1 and λ_2 balance the prior regularization term and reconstruction error term, and KL is the Kullback-Leibler divergence. The reconstruction error term is actually equivalent to the $L1$ norm between x and x', since we assume $p(x|z)$ has a Laplacian distribution.

GAN loss The GAN loss is introduced to improve the photorealistic quality of the generated image. Since the encoder-decoder of VAE is treated as the GAN generator, we use the input image x and generated image $G(x)$ from VAE as the real and fake images for discriminative training. The GAN loss function is as follows:

$$\begin{aligned} \mathcal{L}_{\text{GAN}} &= E_{x \sim p_{\text{data}}(x)}[\log\ D(x)] \\ &+E_{x \sim p_{\text{data}}(x)}[\log\ (1 - D(G(x)))]. \end{aligned} \tag{3}$$

The weights of the generator and discriminator are updated alternatively in the training process.

Cycle consistency loss Other than identity consistence, the consistency of facial characteristics serves as a good constraint for attribute manipulation. Since it is hard to keep track of those charecteristics without supervision, we adopt cycle consistency similar to [20, 42]. Specifically, we impose the cycle consistency constraint along the dimension of attribute. We apply two inverse transformations $G^+(\cdot)$ and $G^-(\cdot)$ with attribute strength $+\delta$ and $-\delta$ to an image x, and ensure the resulting image $G^-(G^+(x))$ resembles the input x. The circle consistency loss is defined as:

$$\mathcal{L}_{\text{Cycle}} = ||G^-(G^+(x_1)) - x_1||_1 + ||G^+(G^-(x_2)) - x_2||_1, \tag{4}$$

where x_1 and x_2 are the training image pair with opposite attribute labels, and we impose the circle consistency constraint for both of them. The $L1$ norm is used to measure the image distance.

ID loss For face attribute manipulation, it is not good enough to make the generated image look photorealistic. Considering an extreme case where one perfectly realistic generated image does not keep any unique traits about the face, it simply does not look alike the original face at all. This is not acceptable for faithful face manipulation. To preserve personal information as much as possible, we use a face recognition network [27] to penalize the shift of face identity, which is one of the most important facial features. Specifically, we extract identify features from

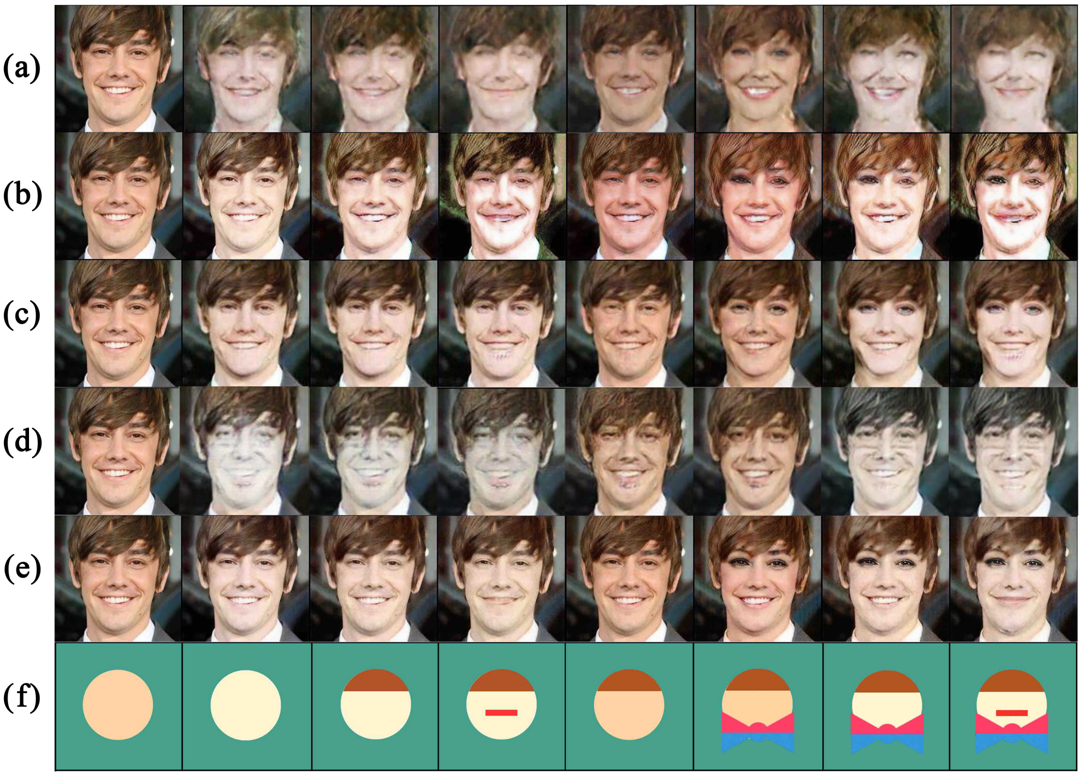

Fig. 3 Facial attribute manipulation results for the 7 typical attributes from CelebA dataset. We compare the state-of-the-art results of (a) residual image GAN, (b) UNIT, (c) StarGAN (d) AttGAN with (e) ours (M-AAE). For each method, the results are shown for the corresponding manipulation for the attributes in the chart (f).

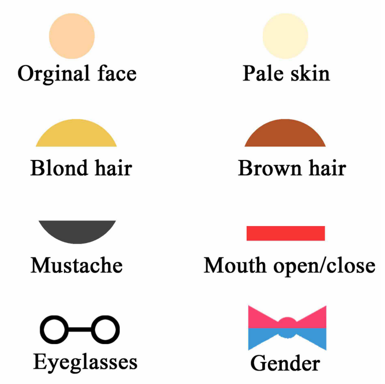

Fig. 4 The manipulated attributes in the paper.

images before and after attribute manipulation, and enforce them to be close to each other. The ID loss function is then defined as:

$$\mathcal{L}_{ID} = \|F_{\text{ID}}(x) - F_{\text{ID}}(G(x))\|^2, \tag{5}$$

where $F_{ID}(\cdot)$ is the feature extractor from the face recognition network.

3.5 Overall Training Procedure

Our final training objective is defined as follows:

$$\begin{aligned} \min_G \max_D \quad & \alpha_1\mathcal{L}_{\text{VAE}} + \alpha_2\mathcal{L}_{\text{GAN}} \\ & + \alpha_3\mathcal{L}_{\text{ID}} + \alpha_4\mathcal{L}_{\text{Cycle}} + \alpha_5\mathcal{L}_{\text{Mask}}, \end{aligned} \tag{6}$$

where the weights of $\alpha_1 \sim \alpha_5$ balance the relative importance of our 5 loss terms. The GAN generator, i.e., the encoder-decoder are trained jointly, while the GAN discriminator is trained alternatively. Further details of the networks may be found in Table 1. The face recognition network is only used to extract features and its weights are frozen. We choose the first 11 layers of the recognition network [27] as feature extractor.

4 Experiments

In this section, we first introduce our used dataset and implementation details. Our M-AAE is compared against state-of-the-arts both qualitatively and quantitatively to show our advantage. Ablation study is conducted to demonstrate the contribution of each component of our framework.

Tab. 2 Image fidelity scores (0 to 1, the higher the better) of different methods for the multi-attribute manipulation task on CelebA dataset.

Num of manipulated attributes	1	2	3	4
Residual image GAN	0.483	0.325	0.330	0.250
UNIT	0.478	0.344	0.374	0.356
StarGAN	0.382	0.344	0.316	0.249
M-AAE(Ours)	**0.521**	**0.507**	**0.398**	**0.365**

Tab. 3 AMT perceptual evaluation for ranking different methods on the multi-attribute manipulation task on CelebA. The average rank (between 1 and 7, from best to worst) is shown in each case. The top cell compares state-of-the-art methods, while the bottom cell compares several baselines of ours.

Num of manipulated attributes	1	2	3	4
Residual image GAN	100%	95.8%	63.9%	33.3%
UNIT	16.7%	87.5%	55.5%	22.9%
StarGAN	33.3%	62.5%	52.3%	75.0%
Modify entire feature map	8.33%	83.3%	47.2%	75.0%
Modify feature map sparsely	100%	91.7%	75.0%	75.0%
ID loss	100%	70.8%	41.7%	62.5%
ID + Mask loss (Ours)	**100%**	**95.8%**	**77.8%**	**77.1%**

4.1 Dataset and Implementation Details

We evaluated on the CelebA dataset [21]. This dataset contains 202599 face images of 10177 celebrities. Images are cropped and re-scaled to 348 × 348 pixels. Each image is labeled with 40 binary attributes, e.g., "hair color", "age", "gender" and "pale skin". We choose 7 typical attributes (see Fig. 3) for our attribute manipulation experiments. For each attribute, we select 1000 test images and train with the remaining images in the dataset.

During training, the face identification network is a model pretrained by using VGG16 which is fixed in the process. Other network weights are initialized from a zero-mean normal distribution with standard deviation 0.02. The learning rate is always set to 0.0001. The loss weights in Eq. (6) are $\alpha_1 = 0.1$, $\alpha_2 = 10$, $\alpha_3 = 20$, $\alpha_4 = \alpha_5 = 80$, and the weights in Eq. (2) are $\lambda_1 = 0.1$, $\lambda_2 = 80$. For training, we use a batch size of 64 and the ADAM [15] optimizer, with a learning rate of 0.0001, betas of 0.5 and 0.999. We treat multi-labels as independent single labels. We separately train one network for each attribute using its available positive-negative sample pairs. During test phase, we sequentially edit multi-labels, i.e., we first change e.g. the hair color using the corresponding network, and then take the generated image as the input to another network that edits the skin color. This avoids enumerating all the label pairs which is almost impossible. For inference, only the generator (encoder-decoder) is used for image generation with varying attributes.

Training process Besides training with the VAE and GAN loss functions, we also use the face recognition loss and cycle consistency loss for faithful preservation of face details. The face recognition module extracts features from images before and after attribute manipulation, and penalizes their feature discrepancy to preserve identity information. While the cycle consistency loss aims to preserve other unique facial information by penalizing the difference between input image x and the generated image after two inverse attribute transformations $G^{+}(x)$ and $G^{-}(x)$. To ensure background consistency, we further generate facial masks to penalize the background difference between input x and output $G(x)$.

Test process We simply feed the input image x through our generator $G(x) = De(En(x))$, changing the relative attribute strength δ in the latent features $En(x)$.

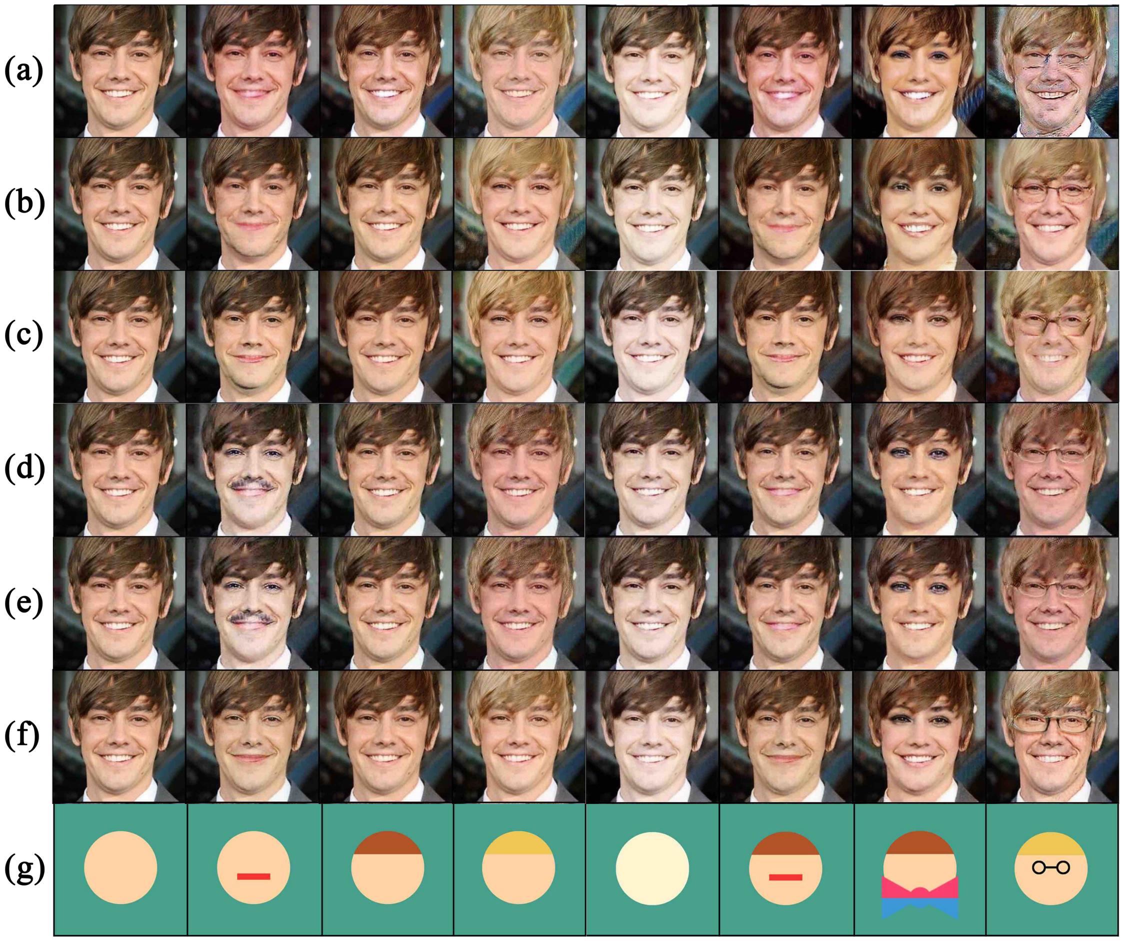

Fig. 5 Comparison of our various baseline in manipulation of the 7 attributes from CelebA dataset. From top to bottom: (a) modify entire feature map, (b) modify feature map sparsely, (c) (b)+ID loss, (d) (c)+Mask loss(concat, raw data), (e) (c)+Mask loss(concat, feature), (f) (c)+Mask loss(Ours). The manipulated attributes for each method are shown in the attribute chart (e).

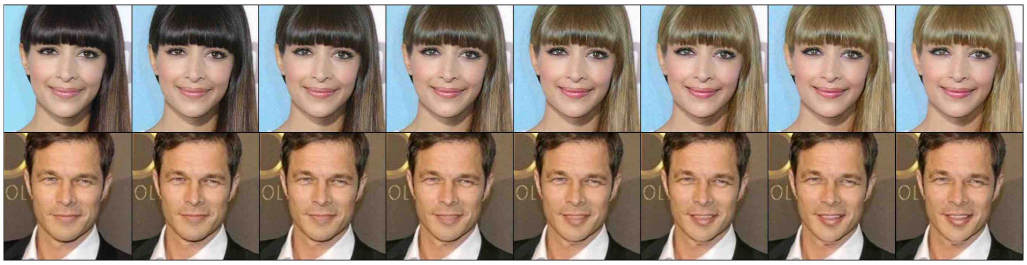

Fig. 6 Continuous manipulation of attributes of blond hair (first row) and mouth open (second row) by our method.

4.2 Qualitative Evaluation

Fig. 3 compares our M-AAE method qualitatively with the state-of-the-art residual image GAN [32], UNIT [20] and StarGAN [4] in the first row. The recent residual image GAN and StarGAN achieve top performance in image translation and attribute manipulation. The UNIT method is similar to ours in using the VAE-GAN framework and cycle-consistency constraint. We observed that all these methods can produce artifacts or lose personal features to some extent. Their performance is usually good on single attribute manipulation or multi-attribute manipulation when the target attributes are correlated (e.g., ”pale skin” and ”gender”). However, the performance deteriorates in more complex scenarios. Especially, residual image GAN totally collapses while generating images with eyeglass. The background generated by previous methods are fuzzy and the color is changed. Easpatially, the residual image GAN and the the UNIT generate the unseen background when we change the eyeglasses attribute. In comparison, our M-AAE method (rightmost, bottom row) consistently produces photorealistic and faithful images with different attributes.

Ablation Study Fig. 5 compares our various baselines to demonstrate the contribution of our major components. From the comparison of results in (a) and (b), we can find that modifying a meaningful subset of feature map pixels can better preserve global face information (e.g., color tone) than modifying the entire feature map. Note the two baselines already use the cycle consistency loss in our VAE-GAN framework, whose efficacy is validated by similar works like UNIT [20]. Hence in (c), we further show that adding an ID loss can enhance the identify preservation while editing other attributes. When we use an extra mask loss in (f), the background is made sharper and the foreground facial details also get enhanced with higher fidelity. From the comparison of results in (d), (e) and (f), our method performs better than concatenation ones by simply modifying a sparse set of feature map pixels.

4.3 Image Fidelity Evaluation

To evaluate the fidelity of our generated face images, we directly use our GAN discriminator to output a fidelity score from 0 to 1. Note the GAN discriminator is trained to distinguish the fake generation from the real, and the higher the fidelity score the better. Table 2 compares the results of state-of-the-art algorithms. It is shown in the table that when the number of changed attribute is increased, the fidelity score is decreased, and the gap between different methods is increasing. The more attributes we change, the more changes the image is get. It is shown that our joint loss can boost the GAN performance, generating images of higher fidelity scores (both on the single and multi-attribute manipulation tasks).

4.4 User Study

We perform a user study by inviting volunteers to evaluate the attribute manipulation results. Given a set of generated images from different methods, the volunteers are instructed to rank the methods based on perceptual realism, quality of transfered attribute and preservation of personal features. The generated images from different methods are shuffled before presented. There are 30 validated volunteers to evaluate results with the 7 attributes chosen from CelebA. The average rank (between 1 and 7, then we convert them to percentage data) of each method is calculated and shown in Table 3. Note that we experiment with different numbers of manipulated attributes from 1 to 4, which have gradually increasing difficulty. The results demonstrate the effectiveness of the proposed method over other alternatives with respect to the rank, especially in the multi-attribute manipulation cases. Our ID loss and Mask loss help improve the results steadily due to their preservation of foreground facial details and background scene.

4.5 Analysis

We show the capability of continuous manipulation of attribute strength in Fig. 6. We achieve this by adjusting the attribute strength between [-5,5] in latent features, which is more favorable than prior methods that take a fixed attribute vector as an input. Moreover, the results in Fig. 7 demonstrate the generalization ability of our method. Our method performs well on the examples with a rich combination of attributes, successfully preserving the unique facial details and background in the generated image with a different attribute.

5 Conclusion and Future Work

In this paper, we propose a Mask-Adversarial AutoEncoder (M-AAE) method to effectively manipulate human face attributes. Our method is the well-extension of the VAE-GAN framework, and we propose an effective method to modify a minimum number of pixels in the feature maps of an encoder, which allows us to change the attribute strength

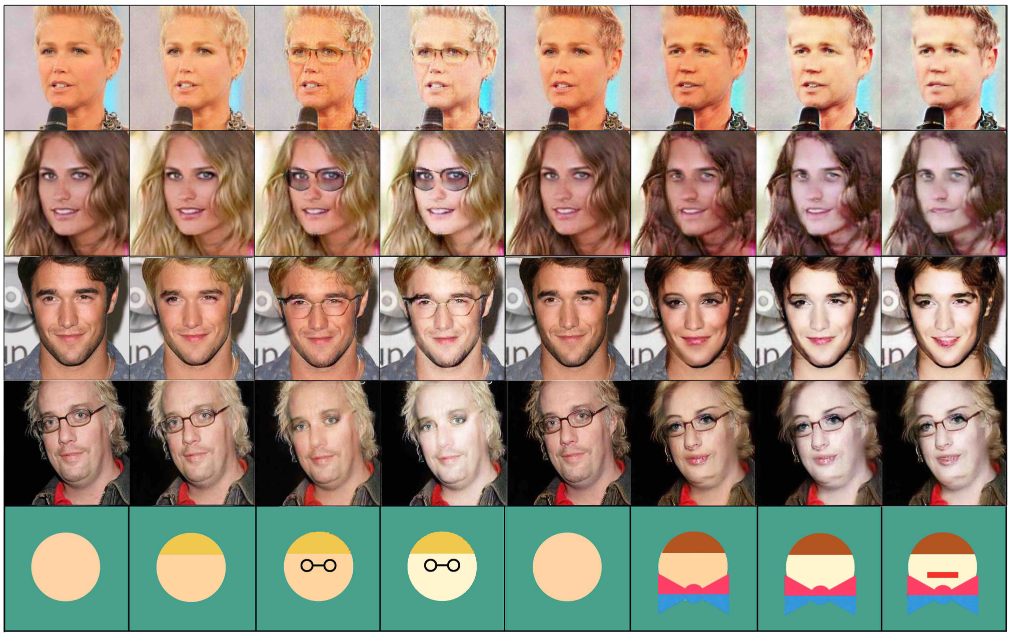

Fig. 7 More results of face attribute manipulation by our M-AAE method. The manipulated attributes for male (first row) are the same as those in Fig. 3, while the manipulated attributes for female (second row) are shown at top-right.

continuously without hindering global information. The proposed network is specifically designed to maintain facial features and image background consistency. We introduce a face recognition loss and a cycle consistency loss for faithful preservation of face details, and also propose a mask loss to ensure background consistency. Experiments show that our method can generate highly photorealistic and faithful images with varying attributes. In principle, our method can be extended to deal with more image translation tasks e.g., style transformation.

Acknowledgements

This paper is partially funded by the National Natural Science Foundation of China (No. 61972157), the National Social Science Foundation of China (No. 18ZD22), the Science and Technology Commission of Shanghai Municipality Program (No. 18D1205903), the Science and Technology Commission of Pudong Municipality Program (No. PKJ2018-Y46), and the Multidisciplinary Project of Shanghai Jiao Tong University (No. ZH2018ZDA25) and is also partially supported by the joint project of SenseTime and Shanghai Jiao Tong University.

References

[1] P. Chen, Q. Xiao, J. Xu, X. Dong, and L. Sun. Facial attribute editing using semantic segmentation. In *2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)*, pages 97–103. IEEE, 2019.

[2] Y.-C. Chen, X. Shen, Z. Lin, X. Lu, I. Pao, J. Jia, et al. Semantic component decomposition for face attribute manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 9859–9867, 2019.

[3] N. D. Chi, K. G. Quach, K. Luu, T. H. N. Le, and M. Savvides. Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3755–3763, 2017.

[4] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *arXiv:1711.09020*, pages 4352–4360, 2017.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. Image

style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[6] J. Gauthier. Conditional generative adversarial nets for convolutional face generation. In *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, volume 2014, pages 2–16, 2014.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[8] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Arbitrary facial attribute editing: Only change what you want. In *arXiv preprint arXiv:1711.10678*, pages 4352–4360, 2017.

[9] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. 2017.

[10] X. Hou, L. Shen, K. Sun, and G. Qiu. Deep feature consistent variational autoencoder. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141. IEEE, 2017.

[11] T. Hung-Yu, L. Hsin-Ying, J. Lu, Y. Ming-Hsuan, and Y. Weilong. Retrievegan: Image synthesis via differentiable patch retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[12] B. Hyojin, C. Sunghyo, Y. Seungjoo, and C. Jaegul. Exploring unlabeled faces for novel attribute discovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 5821–5830, 2020.

[13] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. In *ACM SIGGRAPH*, pages 4352–4360, 2018.

[14] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 4352–4360, 2017.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. pages 4352–4360, 2014.

[16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, pages 4352–4360, 2014.

[17] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4352–4360, 2017.

[18] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning (ICML)*, pages 4352–4360, 2016.

[19] Y. Li, N. Wang, J. Liu, and X. Hou. Demystifying neural style transfer. In *Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2230–2236, 2017.

[20] M. Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 443–449, 2017.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 4352–4360, 2015.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[23] M. X. X. L. E. D. W. Z. S. W. Ming Liu, Yukang Ding. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *CVPR*, pages 3673–3682., 2019.

[24] M. Mirza and S. Osindero. Conditional generative adversarial nets. pages 4352–4360, 2014.

[25] K. Oren, L. Dani, and D. Cohen-Or. Cross-domain cascaded deep translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–6, 2020.

[26] U. Park, Y. Tong, and A. K. Jain. Age-invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32:947–954, 2010.

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference (BMVC)*, volume 41, pages 1–12, 2015.

[28] G. Perarnau, V. D. W. Joost, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4352–4360, 2016.

[29] S. Qian, K. Y. Lin, W. Wu, Y. Liu, Q. Wang, F. Shen, C. Qian, and R. He. Make a face: Towards arbitrary high fidelity face manipulation. In *ICCV*, 2019.

[30] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Computer Science*, pages 4352–4360, 2015.

[31] E. Richardson, M. Sela, R. Orel, and R. Kimmel. Learning detailed face reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5553–5562, 2017.

[32] W. Shen and R. Liu. Learning residual images for face attribute manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 443–449, 2017.

[33] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[34] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and

S. M. Seitz. Total moving face reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 796–812, 2014.

[35] C. Wang, H. Zheng, Z. Yu, Z. Zheng, Z. Gu, and B. Zheng. Discriminative region proposal adversarial networks for high-quality image-to-image translation. In *arXiv preprint arXiv:1711.10678*, pages 4352–4360, 2017.

[36] Z. Wenbin, W. HsiangTao, C. Zeyu, V. Noranart, and W. Baoyuan. Reda:reinforced differentiable attribute for 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2020.

[37] J. Yang, A. Kannan, D. Batra, and D. Parikh. Lr-gan: Layered recursive generative adversarial networks for image generation. In *Conference proceedings: papers accepted to the International Conference on Learning Representations (ICLR)*, pages 4352–4360, 2017.

[38] G. Zhang, M. Kan, S. Shan, and X. Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 417–432, 2018.

[39] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4352–4360, 2017.

[40] Z. Zhang, Y. Song, and H. Qi. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, pages 4352–4360, 2018.

[41] W. Zhou, G. Yang, and S. Hu. Jittor-gan: A fast-training generative adversarial network model zoo based on jittor. In *CVM*, 2021.

[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 4352–4360, 2017.

Ruoqi Sun was born in Weihai, Shandong Province, China, in 1993. She received the B.S. degrees in Digital Media Technology in Shandong University in 2015. And She is currently pursing Ph.D degree with the Department of Computer Science and Engineering in Shanghai Jiao Tong University. Her current research interests include face attribute manipulation, semantic segmenta- tion and image classification.

Chen Huang received the Ph.D. degree in Electronic Engineering from Tsinghua University, Beijing, China, in 2014. He was a postdoctoral fellow in the Robotics Institute of Carnegie Mellon University, and also in the Department of Information Engineering, Chinese University of Hong Kong. He is currently a Research Scientist at Apple Inc. His research interests include machine learning and computer vision, with focus on deep learning and efficient optimization. He has published more than 20 papers in top tier conferences such as CVPR/ICCV/ECCV and NeurIPS/ICML.

Hengliang Zhu received the M.S. degree from the Fujian Normal University, China in 2010. He is now a Ph.D. candidate in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. His current research interests include saliency detection and face alignment.

Lizhuang Ma received the B.S. and Ph.D. de- grees from the Zhejiang University, China in 1985 and 1991, respectively. He is now a Distinguished Professor, Ph.D. Tutor, and the Head of the Digital Media Technology and Data Reconstruction Lab at the Department of Computer Science and En- gineering, Shanghai Jiao Tong University, China. He has published more than 200 academic re- search papers in both domestic and international journals and conferences. His research interests include computer aided geometric design, computer graphics, scientific data visualization, computer animation, digital media technology, and theory and applications for computer graphics, CAD/CAM. He has published more than 200 papers in the famous conference and journals, including more than 100 papers in SCI and EI journals(including IEEE TPAMI, TIP, CVPR, ECCV etc.).