Computational Visual Media
https://doi.org/10.1007/s41095-021-0219-7 Vol. 7, No. 3, September 2021, 363–374
Research Article

Mask-aware photorealistic facial attribute manipulation

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

Abstract The technique of facial attribute manipulation has found increasing application, but it remains challenging to restrict editing of attributes so that a face's unique details are preserved. In this paper, we introduce our method, which we call a *mask-adversarial autoencoder* (M-AAE). It combines a variational autoencoder (VAE) and a generative adversarial network (GAN) for photorealistic image generation. We use partial dilated layers to modify a few pixels in the feature maps of an encoder, changing the attribute strength continuously without hindering global information. Our training objectives for the VAE and GAN are reinforced by supervision of face recognition loss and cycle consistency loss, to faithfully preserve facial details. Moreover, we generate facial masks to enforce background consistency, which allows our training to focus on the foreground face rather than the background. Experimental results demonstrate that our method can generate high-quality images with varying attributes, and outperforms existing methods in detail preservation.

Keywords face attribute manipulation; generative adversarial network (GAN); variational autoencoder (VAE); partial dilated layers; photorealism

1 Introduction

The task of facial attribute manipulation aims to edit facial 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 [1–4]. Essentially, this is an image generation problem. With the advent of generative adversarial networks (GANs), the quality of generated images has improved over time [5, 6]. The family of GAN methods can be mainly divided into two categories: one with noisy input [7, 8], and the other conditioned on input images [9–11]. Our method falls into the second category, aiming to change facial attributes in the input image while preserving high-frequency detail.

Normally, a neural network generates result images by manipulating all pixels of the input image. However, unlike the style translation task [12, 13], the attribute manipulation task 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 such manipulated images in three ways: concentrating attribute manipulation, preserving facial details, and the photorealistic mechanism.

The manipulation method aims to *concentrate attribute manipulation*, i.e., focus on modifying the target attributes while keeping common features unchanged. One simple choice to achieve this goal is to use a conditional GAN framework [7, 14], which concatenates the input image with a one-hot attribute vector encoding the desired manipulation. Another option is to directly learn the image-to-image translation with respect to attributes. CycleGAN [15] learns such a translation rule from unpaired images with a cycle consistency constraint. However, such global transformations can neither guarantee common feature preservation, nor make continuous changes in attribute strength.

Although achieving promising results, the above methods have a common drawback: there is no mechanisms to *preserve facial details*, i.e., to keep unique facial traits while editing whole images. Non-

1 Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: R. Sun, ruoqisun7@sjtu.edu.cn; H. Zhu, hengliang_zhu@sjtu.edu.cn; L. Ma, ma-lz@cs.sjtu.edu.cn (✉).
2 Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA. E-mail: chen-huang@apple.com.
Manuscript received: 2020-12-30; accepted: 2021-02-25

targeted features beyond the background may still be changed, which is undesirable. We especially note the importance of keeping the background unchanged: in practice it is often observed to change along with the foreground face. This suggests some efforts of facial attribute manipulation are wasted on irrelevant regions. Moreover, the post-process of overriding generated background with the original by means of a background mask is undesirable, as it needs careful handling along the boundaries to avoid visible seams.

The *realism* of the generated image is one of the most important aspects of the image generation algorithm, including the fidelity of facial features, the clarity of the image, and so on. Since the features are varied, different methods have been proposed to fit special tasks. The method of Ref. [14] 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. The recent UNIT method [16] uses generative adversarial networks (GANs) and variational autoencoders (VAEs) for robust modeling of different image domains. A cycle consistency constraint is also applied to learn the domain translation effectively. The method of Ref. [17] proposes to only learn the residual image before and after attribute manipulation by using two transformation networks, one for attribute manipulation and the other for the dual operation. However, the methods mentioned above focus on a 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 a partial dilated layer to modify the minimum number of feature map pixels from our encoder. This allows us to maximally preserve global image information and enables attribute change in a continuous manner. Secondly, we feed the background mask into the network to coherently penalize differences before and after facial attribute manipulation. Finally, our method is based on the VAE–GAN framework [14, 16] 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). We call the proposed method a *mask-adversarial autoencoder* (M-AAE). Our experimental results demonstrate its effectiveness.

In summary, the contributions of this paper are: (i) partial dilated layers to modify a few pixels in our learned feature maps, to realize continuous manipulation of facial attributes, (ii) a mask-adversarial autoencoder strategy to ensure faithful facial detail preservation as well as background consistency, and (iii) combining the GAN, VAE, mask loss, ID loss, and 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

2.1 Facial attribute manipulation

Considerable progress has been made in facial attribute manipulation [18–23]. Most methods of facial attribute manipulation are based on generative models. There are two main groups of methods: ones using an extra input vector [9, 14, 24, 25], and the others that directly learn the image-to-image translation along attributes [15, 16]. The first group often takes an attribute vector as guidance for manipulating the desired attribute. The CAAE method [14] concatenates a one-hot age label with latent image features for feeding into the generator for age progression purposes. StarGAN [9] takes a one-hot vector to represent domain information for domain transfer. However, such global transformations based on external codes generally do not preserve facial details well after attribute manipulation. Methods in the second group only operate in image domains and learn the image-to-image translation directly. CycleGAN [15] and the UNIT method [16] are such examples, supervised by a cycle consistency loss that maps the manipulated image back to the original image. Ref. [17] 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 in manipulating the target attribute while keeping others unchanged.

2.2 Image generation algorithm

The VAE [26] and GAN [5] nowadays provide the backbone for image generation tasks such as image

reconstruction [27–30], image synthesis [8, 31, 32], and image translation [33–35]. In a 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 global features extracted from input images. The more recent GAN consists of generator and discriminator networks which play a min–max game. Specifically, the generator tries to synthesize images to fool the discriminator, which in turn distinguishes synthetic images from real ones. GAN-based methods have shown remarkable results in image generation, and many improvements have followed. DCGAN [31] trains stably in a purely convolutional setting, while CGAN [7] generates visually compelling images conditioned on extra input such as class labels. CycleGAN [15] and the UNIT method [16] introduce a cycle consistency loss to learn between image domains with even unpaired images. A recent trend is to combine a GAN with a VAE for robust image modeling. For example, Ref. [19] combines 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 a joint VAE–GAN model is also applied in Refs. [14, 16] for image translation. Recently, high quality generation of human face images uses GANs [36, 37]. Kim et al. [38] proposed utilizing a GAN to transfer a full 3D head expression from a source actor to a target actor in video. We use the VAE–GAN model for facial attribute manipulation, and propose a working method to modify latent VAE features so as to change facial attributes but not irrelevant details.

3 Methodology

3.1 Goals

Our goal is to manipulate the attributes 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 to keep the face faithful: the generated image should look real and its unique details should be preserved, including the background. To address these challenges we propose an M-AAE method.

3.2 Framework overview

Our M-AAE method is based on a VAE–GAN framework, as shown in Fig. 1. The encoder–decoder $De(En(x))$ of the VAE for input image x is treated as the GAN's generator $G(x)$. The discriminator $D(\cdot)$ of the GAN tells the generated image $G(x)$ apart from real images. To manipulate the attributes of the input image x, we use a simple but effective mechanism to uniformly modify the encoded features $En(x)$ by a relative value $\pm\delta$, which is fed into the decoder to control the attribute strength present in the output $G^+(x)/G^-(x)$.

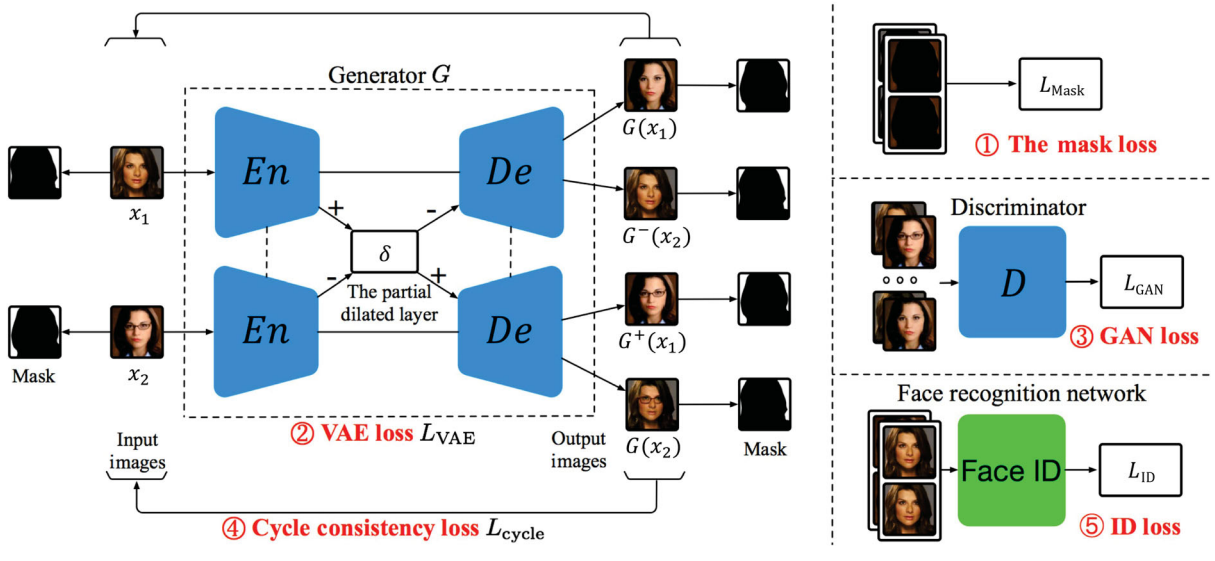


Fig. 1 Framework of our M-AAE method. The encoder–decoder $De(En(x))$ of the VAE for input image x is treated as the generator $G(x)$ of the GAN, with a discriminator $D(\cdot)$ telling fake from real. We manipulate attributes by modifying the encoded features $En(x)$ by a relative value $\pm\delta$,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

We propose partial dilated layers to manipulate the facial features continuously while preserving the consistency of the global features during manipulation. Furthermore, a mask-aware method is utilized to separate the foreground and background of the input image. Thus, the method can focus on the foreground images and manipulate the chosen features only in the foreground, reducing the influence of the background. Modifying image features by using a small number of features instead of modifying all pixels of the image can protect image features. The proposed losses focus on different aspects, including the identity and age of the face, the clarity of the image, and so on. The combination of these losses leads to better image quality.

3.3 Partial dilated layers for attribute manipulation

To manipulate facial attributes, rather than take a one hot attribute vector as in Refs. [9, 14], we choose to modify the hidden features in our encoder: this allows us to continuously change 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 in 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. Figure 2 illustrates how to find such a minimum set of pixels in the partial dilated layer (the last layer of the encoder) recursively from the bottom layer. In this way, image-level manipulation can operate efficiently with modest feature modification. More importantly, we avoid a huge loss of image information. Our experiments show the efficacy of information preservation during attribute manipulation.

In practice, the relative value δ is chosen as half the value range of the feature map pixels to reverse one particular attribute ($\delta \approx 5$ in our scenario). Then the modified features are fed into the decoder to generate an output image $G^+(x)$ or $G^-(x)$ with strengthened or weakened attribute. Adding δ strengthens the facial attribute, and subtracting weakens it. We change the value in the training process (when we apply the cycle consistency loss) to enforce saving strength information in it.

3.4 Mask-aware algorithm for facial detail preservation

In some cases, we observed that the image background would change along with the foreground face using the previous attribute manipulation method. This is not visually pleasing and also suggests some manipulation efforts are wasted in the wrong regions. We claim that pasting the original background around the manipulated face is not ideal: because pixels in the final image coming from images with

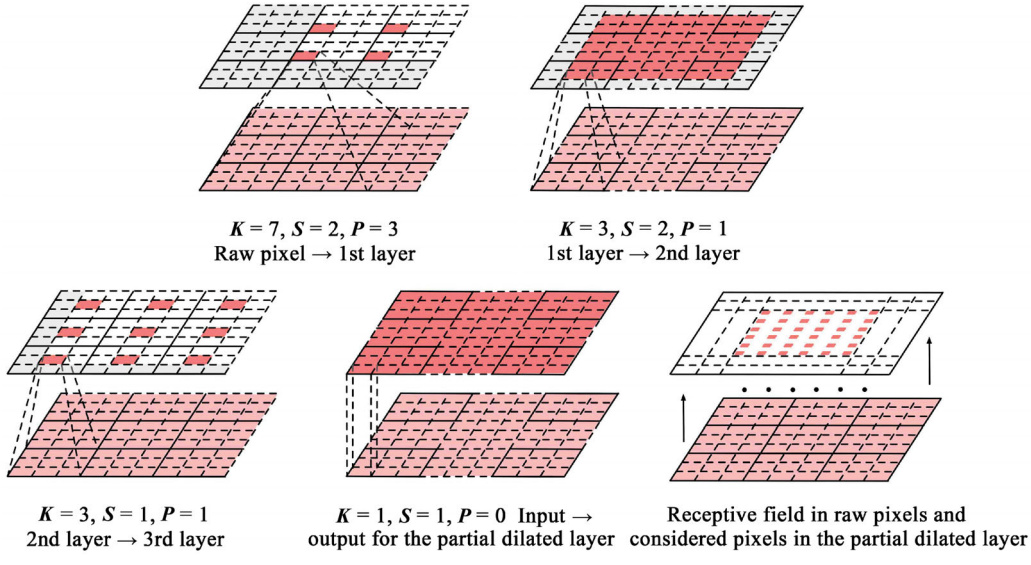


Fig. 2 Receptive fields of the four kinds of partial dilated layers (from bottom to top) of our encoder (K, S, P denote kernel size, stride, and padding, respectively; see Table 1 for details). The padding in each layer during recursive calculation does not belong to the receptive field. At bottom right, the global receptive field in raw pixels and modified pixels in the partial dilated layer are shown. Our goal is to find the minimum number of feature map pixels in the partial dilated layer whose receptive field covers the whole image in the image domain.

different distributions, they seem incompatible. More importantly, it is better to mask out the background at the algorithm level to focus our manipulation efforts on the foreground face. As a side effect, the background remains unchanged. Here we propose a *mask loss* to learn to change the foreground facial attribute and keep the background the same in a coherent way. We generate a facial mask (and thus background mask as well) by using FCN [39], and penalize the background difference between the input x and generated result $G(x)$:

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

where $\text{Mask}(\cdot)$ is the masking operator using the generated background mask. Note that 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 an inconsistent penalty.

3.5 Photorealism mechanism

3.5.1 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. The VAE also penalizes the reconstruction error between x and x', and has loss function:

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

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.

3.5.2 GAN loss

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

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

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

3.5.3 Cycle consistency loss

Other than identity consistency, the consistency of facial characteristics is an important constraint for attribute manipulation. Since it is hard to keep track of those characteristics without supervision, we adopt cycle consistency, following Refs. [15, 16]. 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 cycle consistency loss is defined as

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

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

3.5.4 ID loss

For facial 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 [40] to penalize the shift of face identity, which is one of the most important facial features. Specifically, we extract identify features from images before and after attribute manipulation, and enforce them to be close to each other. The ID loss function is then defined as

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

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

3.6 Overall training procedure

Our overall training objective is defined as

$$\begin{aligned}\min_G \max_D \quad & \alpha_1 \mathcal{L}_{\text{VAE}} + \alpha_2 \mathcal{L}_{\text{GAN}} \\ & + \alpha_3 \mathcal{L}_{\text{ID}} + \alpha_4 \mathcal{L}_{\text{cycle}} + \alpha_5 \mathcal{L}_{\text{mask}} \end{aligned} \quad (6)$$

where the weights of α_1–α_5 balance the relative importance of the 5 loss terms. The encoder–decoder forming the GAN generator are trained jointly, while the GAN discriminator is trained alternately. Further details of the networks may be found in Table 1. The face recognition network is only used to extract

Table 1 Network architecture of our encoder, decoder, and GAN discriminator (number of channels, 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

features and its weights are frozen. We use the first 11 layers of the recognition network [40] as the feature extractor.

4 Experiments

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

4.1 Dataset and implementation details

4.1.1 Dataset

We evaluated methods on the CelebA dataset [41]. It contains 202,599 facial images of 10,177 celebrities. Images are cropped and re-scaled to 348 × 348 pixels. Each image is labeled with 40 binary attributes, including hair color, age, gender, and pale skin. We chose 7 typical attributes (see Fig. 3) for our attribute manipulation experiments. For each attribute, we

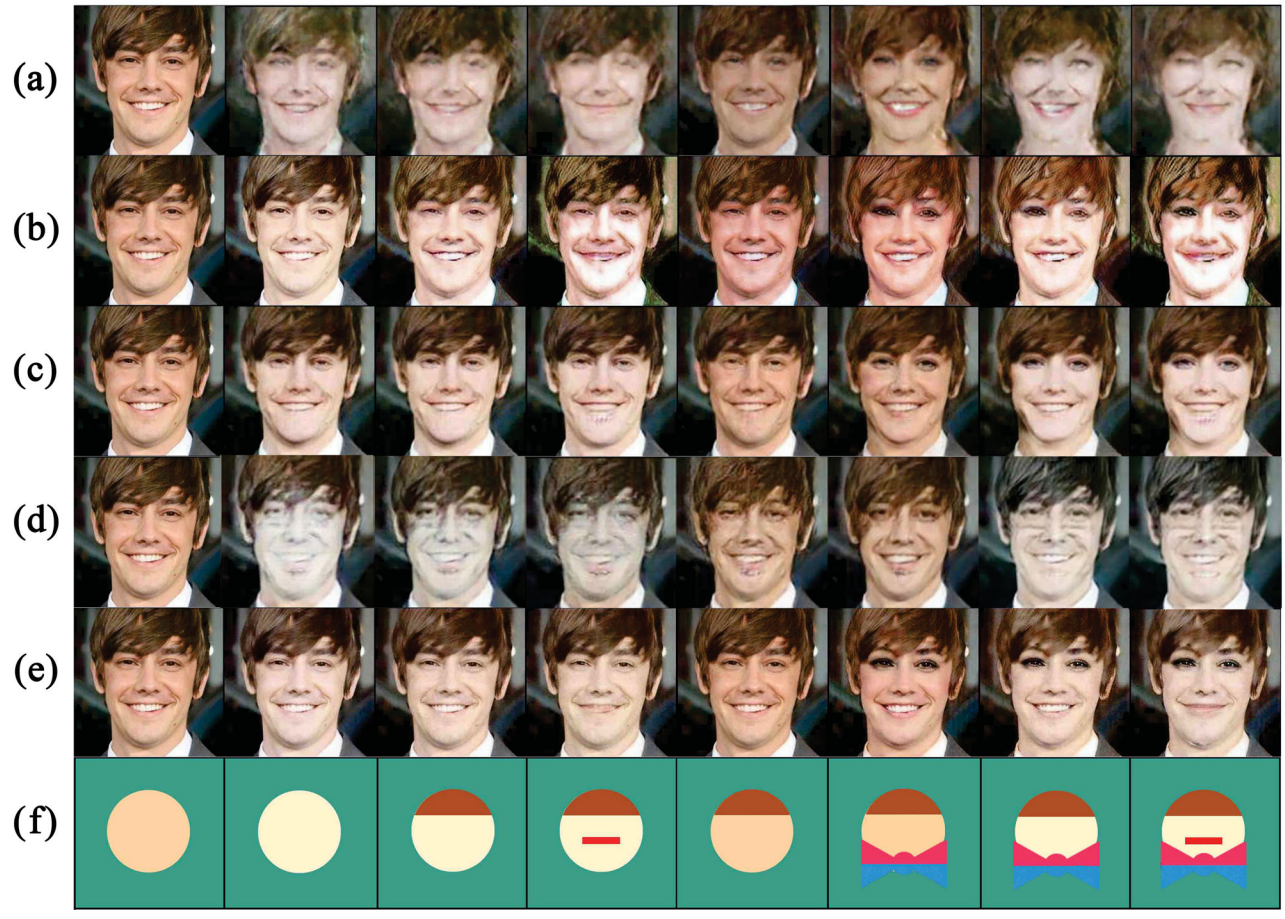


Fig. 3 Facial attribute manipulation results for the 7 typical attributes from the CelebA dataset. We compare the state-of-the-art results of (a) residual image GAN, (b) UNIT, (c) StarGAN, (d) AttGAN with (e) ours (M-AAE). In each case, the attribute manipulated is indicated in (f), with key in Fig. 4.

Fig. 4 Manipulated attributes.

selected 1000 test images and train with the remaining images in the dataset.

4.1.2 Implementation

During training, the face identification network is a model pretrained 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 fixed at 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 [42] 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 testing, we sequentially edit multi-labels, so for example we first change the hair color using the corresponding network, and use the generated result as the input to another network that edits the skin color. This avoids enumerating all label pairs, which is infeasible. In inferencing, only the generator (encoder–decoder) is used for image generation with varying attributes.

4.1.3 Training

Besides training with the VAE and GAN loss functions, we also use the face recognition loss and cycle consistency loss for faithful preservation of facial details. The face recognition module extracts features from images before and after attribute manipulation, and penalizes their feature discrepancy to preserve identity information. The cycle consistency loss aims to preserve other unique facial information by penalizing the difference between the 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 $G(x)$.

4.1.4 Testing

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 $En(x)$.

4.2 Qualitative evaluation

Figure 3 compares our M-AAE method qualitatively with state-of-the-art methods: residual image GAN [17], UNIT [16], and StarGAN [9]. 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 can see that all these methods 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. In particular, residual image GAN totally collapses while generating images with spectacles. The backgrounds generated by previous methods are indistinct and its color is changed. In particular, residual image GAN and UNIT generate an unseen background when we change the spectacles attribute. In comparison, our M-AAE method (rightmost, bottom row) consistently produces photorealistic and faithful images with different attributes.

4.3 Ablation study

Figure 5 compares our various baselines to demonstrate the contribution of our major components. Comparing results in Figs. 5(a) and 5(b), we see 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 in similar work like UNIT [16]. Hence in Fig. 5(c), we further show that adding an ID loss can enhance identify preservation while editing other attributes. When we use an extra mask loss in Fig. 5(f), the background is made sharper and the foreground facial details are also enhanced with higher fidelity. Comparing results in Figs. 5(d)–5(f),

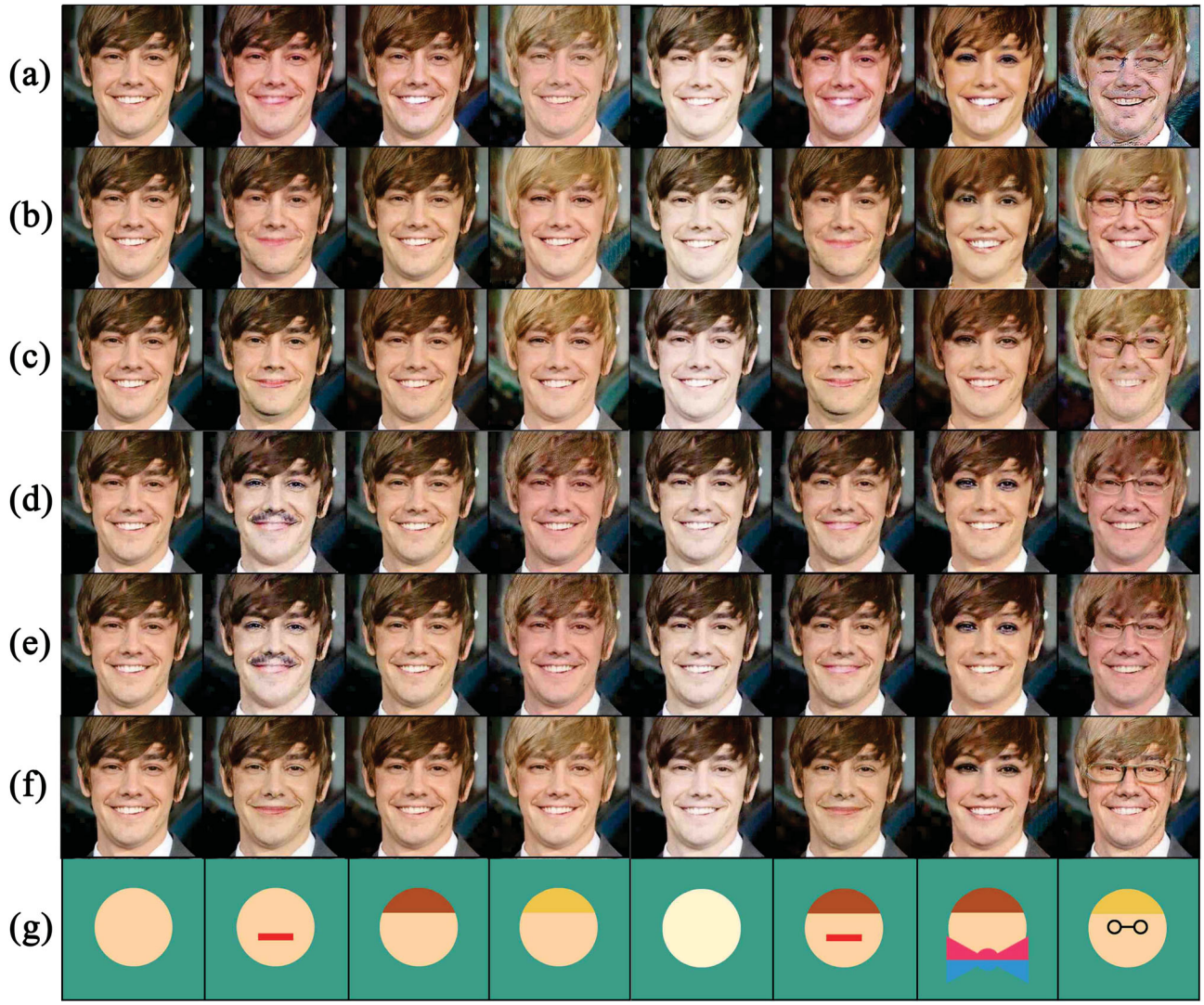


Fig. 5 Comparison of our various baseline in manipulation of the 7 attributes from CelebA dataset. From top to bottom: (a) modify entire feature map, (b) modify feature map sparsely, (c) (b)+ID loss, (d) (c)+mask loss (concat, raw data), (e) (c)+mask loss (concat, feature), (f) (c)+mask loss (ours).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

our method performs better than concatenation ones simply modifying a sparse set of feature map pixels.

4.4 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 generated fakes from real images, and the higher the fidelity score the better. Table 2 compares the results of state-of-the-art algorithms. As the number of changed attribute increases, the fidelity score decreases, and the gap between different methods increases. The more attributes we change, the more changes the image undergoes. Our joint loss boosts GAN performance, generating images with higher fidelity scores, both on single and multi-attribute manipulation tasks.

Table 2 Image fidelity scores, 0 to 1, for different methods for the multi-attribute manipulation task for the CelebA dataset

Number of 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**

4.5 User study

We performed a user study by inviting volunteers to evaluate the attribute manipulation results. Given a set of images generated by different methods,

the volunteers were instructed to rank the methods based on perceptual realism, quality of transferred attribute, and preservation of personal features. The images generated by different methods were shuffled before being presented. 30 volunteers evaluated results with the 7 attributes chosen from CelebA. The average rank (between 1 and 7, then converted to percentages) for each method was calculated and is shown in Table 3. We considered from 1 to 4 manipulated attributes, leading to 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.

4.6 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 flexible than prior methods that take a fixed attribute vector as an input. Moreover, the results in Fig. 7 demonstrate the generalizability of our method. Our method performs well on the examples with a rich combination of attributes, successfully preserving unique facial details and the background in the generated image with a different attribute.

Table 3 Average AMT perceptual evaluation ranking different methods on the multi-attribute manipulation task on CelebA. The top cell compares state-of-the-art methods, while the bottom cell compares several baselines for our method

Num of 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 full feature map	8.33%	83.3%	47.2%	75.0%
Modify part feature map	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%**

5 Conclusions and future work

In this paper, we propose a mask-adversarial autoencoder method to manipulate human facial attributes. Our method extends 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 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 other image translation tasks such as style transformation.

Acknowledgements

This paper was 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 a joint project of SenseTime and Shanghai Jiao Tong University.

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

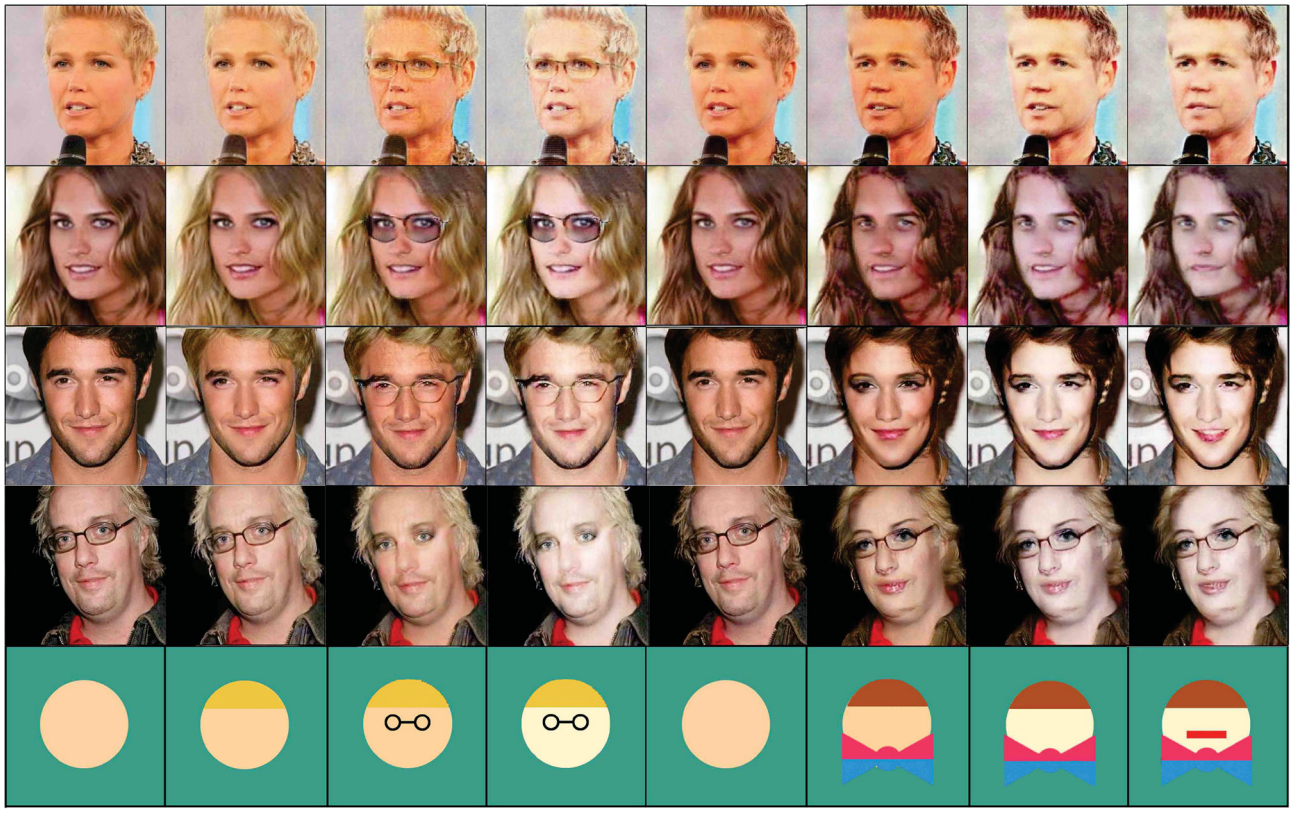


Fig. 7 Further facial attribute manipulation results using our M-AAE method. The manipulated attributes for male (first row) are the same as those in Fig. 3, while the manipulated attributes for female (second row) are shown at top-right.

References

[1] Park, U.; Tong, Y. Y.; Jain, A. K. Age-invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 32, No. 5, 947–954, 2010.

[2] Duong, C. N.; Quach, K. G.; Luu, K.; Le, T. H. N.; Savvides, M. Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 3755–3763, 2017.

[3] Zhang, G.; Kan, M. N.; Shan, S. G.; Chen, X. L. Generative adversarial network with spatial attention for face attribute editing. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11210.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 422–437, 2018.

[4] Qian, S.; Lin, K.; Wu, W.; Liu, Y.; Wang, Q.; Shen, F.; Qian, C.; He, R. Make a face: Towards arbitrary high fidelity face manipulation. In: Proceedings of the International Conference on Computer Vision, 10033–10042, 2019.

[5] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y.; Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, Vol. 2, 2672–2680, 2014.

[6] Zhou, W. Y.; Yang, G. W.; Hu, S. M. Jittor-GAN: A fast-training generative adversarial network model zoo based on Jittor. *Computational Visual Media* Vol. 7, No. 1, 153–157, 2021.

[7] Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv preprint* arXiv:1411.1784, 2014.

[8] Yang, J.; Kannan, A.; Batra, D.; Parikh, D. LRGAN: Layered recursive generative adversarial networks forimage generation. In: Proceedings of the International Conference on Learning Representations, 2017.

[9] Choi, Y.; Choi, M.; Kim, M.; Ha, J. W.; Choo, J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8789–8797, 2018.

[10] Chen, Y. C.; Shen, X. H.; Lin, Z.; Lu, X.; Pao, I. M.; Jia, J. Y. Semantic component decomposition for face attribute manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9851–9859, 2019.

[11] Liu, M.; Ding, Y.; Xia, M.; Liu, X.; Ding, E.; Zuo, W.; Wen, S. STGAN: A unified selective transfer network

for arbitrary image attribute editing. In: Proceedings of the Computer Vision and Pattern Recognition, 3673–3682, 2019.

[12] Gatys, L. A.; Ecker, A. S.; Bethge, M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2414–2423, 2016.

[13] Li, Y. H.; Wang, N. Y.; Liu, J. Y.; Hou, X. D. Demystifying neural style transfer. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2230–2236, 2017.

[14] Zhang, Z. F.; Song, Y.; Qi, H. R. Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5810–5818 2017.

[15] Zhu, J. Y.; Park, T.; Isola, P.; Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2242–2251, 2017.

[16] Liu, M. Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. In: Proceedings of the 31st Conference on Neural Information Processing Systems, 443–449, 2017.

[17] Shen, W.; Liu, R. J. Learning residual images for face attribute manipulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1225–1233, 2017.

[18] Lample, G.; Zeghidour, N.; Usunier, N.; Bordes, A.; Denoyer, L.; Ranzato, M. Fader networks: Manipulating images by sliding attributes. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 5969–5978, 2017.

[19] Larsen, A.; Sønderby S.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. In: Proceedings of the International Conference on Machine Learning, 1558–1566, 2016.

[20] He, Z.; Zuo, W.; Kan, M.; Shan, S.; Chen, X Arbitrary facial attribute editing: Only change what you want. *arXiv preprint* arXiv:1711.10678, 2017.

[21] He, Z.; Zuo, W.; Kan, M.; Shan, S.; Chen, X. AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing* Vol. 28, No. 11, 5464–5478, 2019.

[22] Chen, P.; Xiao, Q.; Xu, J.; Dong, X. L.; Sun, L. J. Facial attribute editing using semantic segmentation. In: Proceedings of the International Conference on High Performance Big Data and Intelligent Systems, 97–103, 2019.

[23] Bahng, H.; Chung, S.; Yoo, S.; Choo, J. Exploring unlabeled faces for novel attribute discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5820–5829, 2020.

[24] Gauthier, J. Conditional generative adversarial nets for convolutional face generation. 2014. Available at http://cs231n.stanford.edu/reports/2015/pdfs/jgauthie_final_report.pdf.

[25] Perarnau, G.; Joost, V.; Raducanu, B.; Alvarez, J. Invertible conditional GANs for image editing. In: Proceedings of the Advances in Neural Information Processing Systems, 2016.

[26] Kingma, D. P.; Welling, M. Auto-encoding variational Bayes. In: Proceedings of the International Conference on Learning Representations, 2014.

[27] Suwajanakorn, S.; Kemelmacher-Shlizerman, I.; Seitz, S. M. Total moving face reconstruction. In: *Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Vol. 8692*. Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 796–812, 2014.

[28] Hou, X. X.; Shen, L. L.; Sun, K.; Qiu, G. P. Deep feature consistent variational autoencoder. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 1133–1141, 2017.

[29] Richardson, E.; Sela, M. T.; Or-El, R.; Kimmel, R. Learning detailed face reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5553–5562, 2017.

[30] Zhu, W. B.; Wu, H. T.; Chen, Z. Y.; Vesdapunt, N.; Wang, B. Y. ReDA: Reinforced differentiable attribute for 3D face reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4957–4966, 2020.

[31] Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint* arXiv:1511.06434, 2016.

[32] Tseng, H. Y.; Lee, H. Y.; Jiang, L.; Yang, M. H.; Yang, W. L. RetrieveGAN: Image synthesis via differentiable patch retrieval. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12353*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 242–257, 2020.

[33] Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, Vol. 70, 1857–1865, 2017.

[34] Shen, Y. J.; Gu, J. J.; Tang, X. O.; Zhou, B. L. Interpreting the latent space of GANs for semantic face editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9240–9249, 2020.

[35] Oren, K.; Dani, L.; Cohen-Or, D. Cross-domain cascaded deep translation. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12347.*

Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 673–689, 2020.

[36] Zhang, Z.; Song, Y.; Qi, H. Progressive growing of GANs for improved quality, stability, and variation. In: Proceedings of the International Conference on Learning Representations, 2018.

[37] Wang, C.; Zheng, H. Y.; Yu, Z. B.; Zheng, Z. Q.; Gu, Z. R.; Zheng, B. Discriminative region proposal adversarial networks for high-quality image-to-image translation. In: Proceedings of the European Conference on Computer Vision, 770–785, 2018.

[38] Kim, H.; Garrido, P.; Tewari, A.; Xu, W. P.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; Theobalt, C. Deep video portraits. *ACM Transactions on Graphics* Vol. 37, No. 4, Article No. 163, 2018.

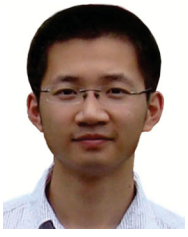
[39] Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440, 2015.

[40] Parkhi, O. M.; Vedaldi, A.; Zisserman, A. Deep face recognition. In: Proceedings of the British Machine Vision Conference, 41.1–41.12, 2015.

[41] Liu, Z. W.; Luo, P.; Wang, X. G.; Tang, X. O. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, 3730–3738, 2015.

[42] Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations, 2015.

Ruoqi Sun was born in Weihai, Shandong Province, China, in 1993. She received her B.S. degree in digital media technology from Shandong University in 2015. She is currently pursuing a Ph.D. degree in the Department of Computer Science and Engineering in Shanghai Jiao Tong University. Her current research interests include facial attribute manipulation, semantic segmentation, and image classification.



Chen Huang received his 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, the Chinese University of Hong Kong. He is currently a Research Scientist at Apple Inc. His research interests include machine learning and computer vision, with a focus on deep learning and efficient optimization. He has published more than 20 papers in top tier conferences such as CVPR, ICCV, ECCV, NeurIPS, and ICML.

Hengliang Zhu received his M.S. degree from 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. His current research interests include saliency detection and face alignment.

Lizhuang Ma received his B.S. and Ph.D. degrees from Zhejiang University, China, in 1985 and 1991, respectively. He is now a Distinguished Professor and Head of the Digital Media Technology and Data Reconstruction Lab at the Department of Computer Science and Engineering, Shanghai Jiao Tong University. He has published more than 200 academic research papers. His research interests include computer aided geometric design, computer graphics, scientific data visualization, computer animation, digital media technology, and theory and applications of computer graphics and CAD/CAM.