Attributes Aware Face Generation with Generative Adversarial Networks

Zheng Yuan*†, Jie Zhang*, Shiguang Shan*†, Xilin Chen*†
*Institute of Computing Technology, Chinese Academy of Sciences, China
†University of Chinese Academy of Sciences, Beijing, China

***Abstract*—Recent studies have shown remarkable success in face image generations. However, most of the existing methods only generate face images from random noise, and cannot generate face images according to the specific attributes. In this paper, we focus on the problem of face synthesis from attributes, which aims at generating faces with specific characteristics corresponding to the given attributes. To this end, we propose a novel attributes aware face image generator method with generative adversarial networks called AFGAN. Specifically, we firstly propose a two-path embedding layer and self-attention mechanism to convert binary attribute vector to rich attribute features. Then three stacked generators generate 64×64, 128×128 and 256×256 resolution face images respectively by taking the attribute features as input. In addition, an image-attribute matching loss is proposed to enhance the correlation between the generated images and input attributes. Extensive experiments on CelebA demonstrate the superiority of our AFGAN in terms of both qualitative and quantitative evaluations.**

I. Introduction

Recently, more and more attention has been paid to image generation. Great progresses have been achieved by generative adversarial networks and its variants [2], [3]. Different from the general image generation, the face synthesis pays more attention on the details of the generated face. Our work aims at generating face images with specific input attributes, i.e., attributes aware face generation, which has not received a wide range of attention in the past. Face generation with specific attributes has wide application prospects. For example, it can be utilized to extend face datasets for improving the face recognition models or provide a face synthesis of the suspect for criminal investigations.

Nowadays, there are several breakthroughs both in diversity and clarity for face generation. The StyleGAN [4] and its upgraded version StyleGANv2 [5], as one of the state-of-the-art methods in face generation, can generate a high resolution face image of 1024×1024 pixels. The generated high resolution image contains detailed face information, e.g., the hairstyle, the thickness of eyebrows, the beard type, etc. However, the attributes of faces are randomly generated and a face with some specific attributes can not be achieved.

Text-to-image generation is another work related to attributes aware face generation. AttnGAN [1], as one of state-of-the-art methods, generates images through a stacked generator and multimodal similarity module. The image generated by AttnGAN is closely related to the input text information and can well characterize the scene described by the text content. However the input attribute labels are different from the text sentence, it is non-trivial to directly apply AttnGAN for attributes aware face generation. DM-GAN [6], Obj-GAN [7] and OP-GAN [8] are recently proposed to further solve the task of text-to-image generation. Due to the dynamic memory mechanism and image layout information used in these works, the generated image is more consistent with the input text.

In the task of attribute-driven face image generation, which are closely related to our work, [9], [10], [11] are proposed to solve this problem, but the resolution and clarity of the generated images still need to be improved.

In this paper, we propose a novel framework named Attributes Aware Face Generation with Generative Adversarial Networks (AFGAN) for generating face images from the input attributes. The model is mainly composed of three modules: attribute embedding module, stacked image generation module and similarity constrain module. In attribute embedding module, we propose a two-path embedding layer to convert the input attribute vector into rich face attribute feature, which is different from the naive embedding table in the word embedding layer from AttnGAN [1]. Since in our task, whether the value of the attribute is 0 or 1, the input of face attribute always has certain meaning. Such as ”young” attribute, when the value of this attribute is 1, it means young people and when the value is 0, it indicates old people. So we design a two-path embedding layer to make the input attributes well reflect their meanings, which benefits to the subsequent image generation. At the same time, we also use the self-attention module [12] right after the embedding layer to introduce interconnections between different attributes. In stacked image generation module, we use a stacked three-level image generator to generate images of 64×64, 128×128 and 256×256 pixels respectively. On the one hand, the coarse-to-fine framework of image generation can reduce the learning burden of image generator. On the other hand, it can gradually improve the quality of the generated image, so as to we achieve a clear and realistic image which is consistent with the input face attributes. In the similarity constrain module, the high-level face feature vector is firstly extracted from the 256×256 face image through a pre-trained Inception_v3 model of face attributes prediction. Then the high-level face feature vector and the face attribute feature from the two-path embedding layer are transformed into the same feature space. By minimizing the distance between the two feature vectors, the generated face image can well reflect the input face attributes.

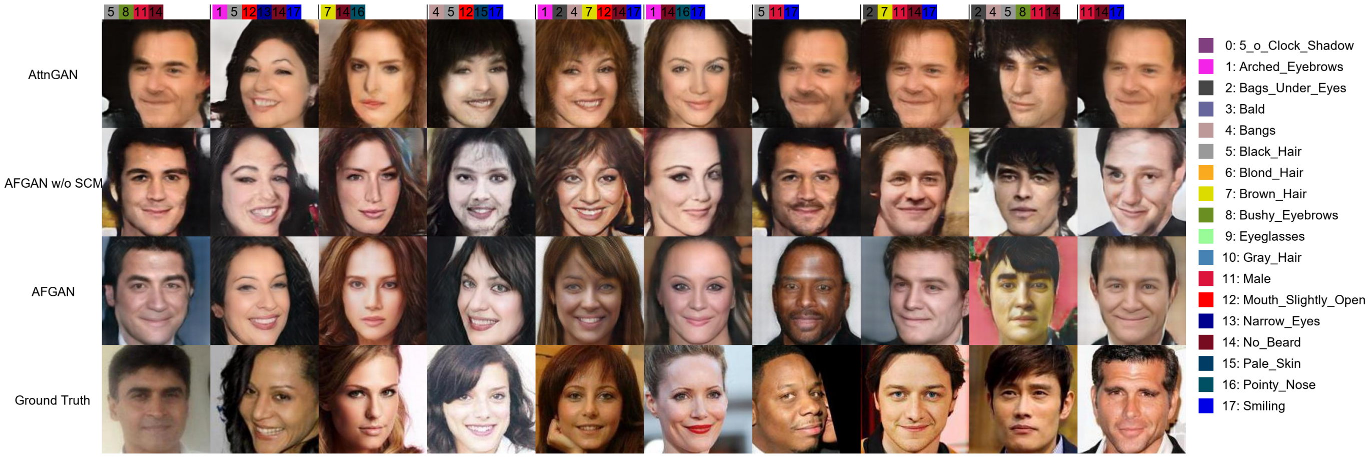

Fig. 1. The generated face images in ablation study of AFGAN model. The first row represents the results of AttnGAN [1], second row represents the experiment results without the constraints of SCM module in the objective function, and third row represents the results of the complete AFGAN model proposed by us. The input attribute label is shown in the top color block and number, and the corresponding attribute content is listed on the right side. The ground truth images corresponding to the attribute labels are shown in the last row.

The contributions of our work are summarized as follow:

- We propose a novel framework named AFGAN for generating face images from face attributes.
- A two-path embedding layer of face attributes is proposed to well characterize the attributes for face generation. The attributes information can be accurately transmitted to the subsequent image generation whether the value of each face attribute is 0 or 1.
- Both qualitative and quantitative experiment results show that the face images generated by our AFGAN can not only conform to the input face attributes, but also have facial details with good image quality and clarity.

II. Related Work

Our work is aimed at generating face images according to the input attributes. We expect that the generated images can fully reflect the characteristic of the corresponding attributes. Limited to the scope of our study, the most related works are high-resolution face image synthesis, text-to-image generations, face hallucination and attribute-driven face image synthesis.

A. *High-resolution face image synthesis*

Face image synthesis, as one typical topic of image synthesis, has received wide attentions since the proposal of generative adversarial networks [2]. Different from the conventional image generation task, which pays more attention to the general structure and the shape of object in the generated image, the task of face image synthesis focuses on the details of the face, such as the texture of the hair, facial wrinkles and skin gloss, etc. Recently there have been several state-of-the-art methods [13], [4], which focus on generating high-resolution realistic faces. PGGAN [13] proposes a novel framework which gradually upsamples the generated images from 4×4 to 1024×1024. The network structure of model is gradually deepened, starting at a resolution of 4×4, and gradually double the resolution of the image by adding an upsampling module. Finally PGGAN can generate high-quality faces with 1024×1024 pixels. StyleGAN [4], as one state-of-the-art work of face generation, adds a hidden vector mapping module based on PGGAN, which embeds the original noise to hidden vectors via eight fully connected layers. Then the generated hidden vectors utilized as AdaIN [14] factor are fed to each layer for different resolution image generation. At the same time, StyleGAN also introduces noises into each layer to increase the diversity of the generated face images from coarse to fine. Although the faces generated from these works are realistic and some of them can mix the spurious with the genuine, we can not achieve faces according to specific attributes.

B. *Text-to-image generation*

Text-to-image generation takes the text description as input and generates the corresponding image which is consistent with the semantic of text. AttnGAN [1] uses a stacked three-stage structure to generate images from 64×64 to 256×256 resolution gradually. And a DAMSM module is proposed to constrain the distance between the text and generated image by the method of encoding image and text to the same semantic space with attention mechanism [12]. MirrorGAN [15] draws on the ideas of AttnGAN [1] and CycleGAN [16] at the same time. The generated image is encoded to text again and constrains the generated text as similar as the input text.

The methods mentioned above can well generate images containing some objects and attributes described by text, but the generated images are mainly focused on general objects like flowers and birds rather than faces. The face generation should pay more attention to the details of the texture of the hair, facial wrinkles and skin gloss, etc., which may be more difficult to generate than general objects. Moreover, generation by attribute labels is also different from that by text descriptions. Our proposed AFGAN takes the attribute vectors as input instead of the text descriptions, which focuses on the interrelations between different attributes. Our method can well generate faces with specific attributes which ensures both the clarity and authenticity.

C. Face hallucination

Some works of face hallucination are also related to attribute-guided face image synthesis. Yu et al. [17] uses an attribute-embeded upsampling network to generate high resolution images from tiny unaligned face images, which reducing the uncertainty of one-to-many mappings remarkably. Li et. al. [18] constructs a face transfer network, which upsamples low-resolution face images to high-resolution images by fusing facial attributes. The difference between face hallucination and our attribute-guided face images synthesis is that our approach has no input images as reference.

D. Attribute-driven face image synthesis

There are also some works focusing on attribute-driven face image synthesis, and we give discussions on differences between them and ours.

Attribute2sketch2face [19] firstly synthesizes the facial sketch corresponding to the input attributes and then reconstructs the face image based on the synthesized sketch. However, the resolution of the generated face image is limited to 64×64. And this work needs the facial sketch as the input. Differently our method designs a stacked three-stage generator to achieve a high resolution facial image by only taking the attribute label as input.

For Lu et al. [9], two variants of CycleGAN are proposed to generate attribute-guided images and identity-guided images respectively. This method takes low resolution faces as input, and combines attributes to generate high-resolution face images. Differently, our AFGAN does not utilize any faces as input, but purely generates face images from attributes.

Yan et al. [10] considers that an image is a combination of foreground and background. A VAE framework is employed to generate face image through disentangled latent variables. However, this work only generates 64 × 64 images, and our work focuses on how to generate higher resolution images, i.e., 256 × 256. Besides, the image generated by Yan et al. [10] looks fuzzy and lacks of diversity while our AFGAN can generates higher quality images.

For Wang et al. [11], a DCGAN-based model is proposed to generate face images by adding attribute vector to the input. At the same time, generation of continuous sequence of face images is also considered in this work. However, the generated images shown in [11] are fuzzy, and our method outperforms [11] in terms of both IS and FID metrics. Moreover, Wang et al. [11] only focuses on five basic attributes, e.g. glasses, gender, hair color, smile and age, while our work focuses on 18 attributes, including more attributes with some particular ones, like pointy nose, bushy eyebrows and so on.

However, the resolution of images generated by these works ranges from 64 to 128, and the image size of our proposed method is 256. So the quality and clarity of the images generated by our method are much better than these works.

III. Approach

In this section, we will start with the overview of our method, and then introduce the three modules respectively. Finally we give the objective function of the whole model.

A. Overview

As shown in Figure 2, AFGAN consists of three modules: attribute embedding module (AEM), stacked image generation module (SIGM) and similarity constrain module (SCM). AFGAN takes the attribute vector as input. In the AEM, the attribute vector is converted to global attribute features and local attribute features through the two-path embedding layer and self-attention layer. In the SIGM, the global attribute feature firstly uses conditioning augmentation method [20] to increase the diversity of the input attribute. Then it is utilized to generate face images through stacked image generator, from low resolution to high resolution gradually together with the local attribute feature. In the SCM, we first encode the generated images through a pretrained inception_v3 [21] model to extract high-level features. Then we convert the encoded image feature and local attribute feature into the same semantic space. Through constraining the distance of two features by extra objective function term, we aim at forcing the generated face images to keep pace with the input face attribute.

B. AEM: Attribute Embedding Module

We denote the input attribute vector as $S_{attr} \in \{0,1\}^N$, N is the number of input attributes. We first convert it into $S_{global} \in \mathbb{R}^C$, which represents the global semantics of attribute vectors and $S_{local} \in \mathbb{R}^{N\times C}$, which represents the semantics of each attribute. C is the dimension of feature space. Through different levels of semantic vectors, we hope that both the high-level semantic information and low-level detailed information existing in the input attribute label can be transmitted during the process of image generation. Specifically, S_{global} is transformed through a fully connected layer:

$$S_{global} = W_{global} * S_{attr}, \tag{1}$$

where $W_{global} \in \mathbb{R}^{C\times N}$. S_{local} is transformed by our proposed two-path embedding layer, as shown in Figure 3, and the process can be formulated as the following:

$$S_{local} = embed_1 * S_{attr} + embed_2 * (1 - S_{attr}), \tag{2}$$

where $embed_1, embed_2 \in \mathbb{R}^{C\times N}$ are two embedding tables, representing the semantics of each input attribute whose value is 0 and 1 respectively. The reason behind this design is that in our work, whether the value of the attribute is 0 or 1, the input of face attribute always has certain meaning. Such as ”young” attribute, when the value of this attribute is 1, it means young people while 0 indicates old people. The carefully designed two-path embedding layer can generate feature vector which well reflects their meanings of the input attribute.

Next, we use self-attention layer [12] to focus on modeling the relationships between different attributes:

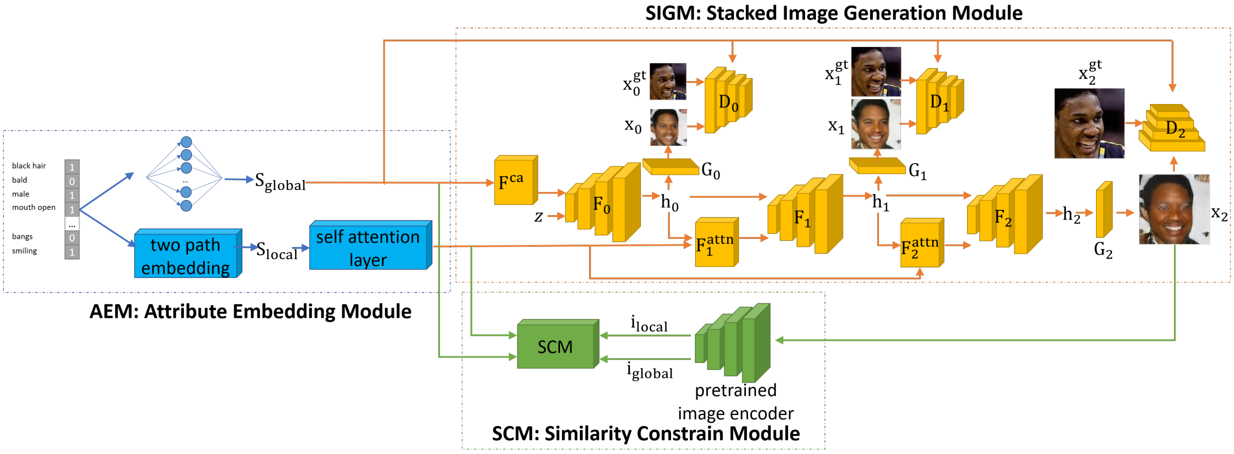

Fig. 2. The structure of AFGAN. The model consists of three modules: attribute embedding module (AEM), stacked image generation module (SIGM) and similarity constrain module (SCM). In AEM, S_{global} and S_{local} represent the global and local features of input attributes. In SIGM, z is the noise vector. F^{ca} denotes conditioning augmentation module [20]. F_i^{attn}, F_i, G_i and D_i are attention module, upsampling block, generation module and discriminator in each stage respectively. h_i is the hidden state vector transmitted in different stages. x_i and x_i^{gt} are the generated and corresponding ground truth images in each stage. In SCM, i_{local} and i_{global} represent the local and global image features extracted by a pretrained image encoder.

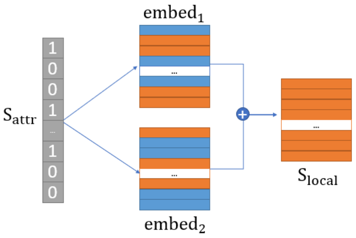

Fig. 3. The structure of two path embedding layer. S_{attr} is the one-hot attribute vector. $embed_1$ and $embed_2$ are two embedding tables, representing the semantics of each input attribute whose value is 0 and 1 respectively. S_{local} represents the semantics of each attribute. The orange vectors in the figure mean the activated vectors.

$$f(x) = W_f * S_{local}, \quad g(x) = W_g * S_{local}, \tag{3}$$

$$s_{ij} = f(x_i)^T g(x_j), \quad \beta_{i,j} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}, \tag{4}$$

$$h(x) = W_h * S_{local}, \quad S_{local'_j} = \sum_{i=1}^{N} \beta_{i,j} h(x_i), \tag{5}$$

$$S_{local'} = \left(S_{local'_1}, S_{local'_2}, \ldots, S_{local'_N}\right) \in \mathbb{R}^{C \times N}, \tag{6}$$

where W_f and W_g are query and key matrix respectively, used to calculate the correlation between the different attributes, and W_h is value matrix. Through self-attention layer, we can achieve enhanced attribute features $S_{local'}$ which taking into account the relationship between different attributes.

C. SIGM: Stacked Image Generation Module

Stacked image generation module is composed of three stacked image generation modules. Firstly, an image of 64 × 64 pixels is generated, and then the scale of the generated image is gradually increased to 128 × 128. Finally the 128 × 128 pixels image is transitioned to a high resolution version of 256 × 256 pixels. The three generation modules are respectively recorded as G_0, G_1 and G_2, and their corresponding hidden state vectors transmitted between different generations are recorded as h_0, h_1, h_2. The corresponding generated images are recorded as x_0, x_1, x_2. The whole process can be formulated as follow:

$$h_0 = F_0\left(z, F^{ca}(S_{global})\right), \tag{7}$$

$$h_i = F_i\left(h_{i-1}, F_i^{attn}(S_{local'}, h_{i-1})\right) \text{ for } i = 1, 2, \tag{8}$$

$$x_i = G_i(h_i), \tag{9}$$

where z is the noise vector sampled from the standard normal distribution, F^{ca} denotes conditioning augmentation module [20] used for converting global attribute vector S_{global} to diverse conditional vector, F_i^{attn} is the attention module used in stage i.

Attention module F^{attn} is utilized to calculate the attribute-content matrix in each stage of image generation, which represents the relationship between each attribute of the input and each region in the generated image. F^{attn} has two inputs: local attribute feature $S_{local'}$ and the hidden state vector from the previous layer. The attribute-content matrix is calculated as follow:

$$s'_{j,i} = h_j^T S_{local_i}, \quad \beta_{j,i} = \frac{\exp\left(s'_{j,i}\right)}{\sum_{k=1}^{N} \exp\left(s'_{j,k}\right)}, \tag{10}$$

$$F^{attn}(S_{local}, h) = (c_1, c_1, \ldots, c_N) \in \mathbb{R}^{D \times N}, \quad c_j = \sum_{i=1}^{N} \beta_{j,i} S_{local_i}. \tag{11}$$

At the same time, each generation module corresponds to a discriminator module D_i, which is used to judge the quality of the generated image. Each discriminator consists of an unconditional discriminator and a conditional discriminator. The input of the former is only an image, which focusing on the authenticity of the generated image, and the inputs of the latter are both the image and the corresponding attribute vector. It is used to judge the matching degree between the generated image and the input attribute.

D. SCM: Similarity Constrain Module

Similarity constrain module is used to constrain the generated images to be more consistent with the input attributes.

We first encode the generated images with a pretrained inception_v3 [21] model, which is trained as an attribute predictor, to achieve high level feature of the image. Specifically, we use the feature from "mixed_6e" layer as local image feature $i_{local} \in \mathbb{R}^{768\times289}$, where 768 is the feature dimension and $289(17*17)$ denotes the number of regions in the image. And we use the feature from the last global pooling layer as the global image feature $i_{global} \in \mathbb{R}^{2048}$.

We use the local attribute feature $S_{local'}$ and local image feature i_{local} to calculate the matching degree between the generated image and input attribute in terms of local evaluation:

$$s = S_{local'}^{T} i_{local} \in \mathbb{R}^{N\times289}, \quad \bar{s}_{i,j} = \frac{\exp\left(s_{i,j}\right)}{\sum_{k=1}^{N}\exp\left(s_{k,j}\right)}, \tag{12}$$

$$\alpha_{i,j} = \frac{\exp\left(\gamma_1 \bar{s}_{i,j}\right)}{\sum_{k=1}^{289}\exp\left(\gamma_1 \bar{s}_{i,k}\right)}, \quad c_i = \sum_{j=1}^{289} \alpha_{i,j} i_{local_j}, \tag{13}$$

$$R\left(c_i, S_{local'_i}\right) = \frac{c_i^T S_{local'_i}}{\|c_i\| \left\|S_{local'_i}\right\|}, \tag{14}$$

$$R^{local}(Q,D) = \log\left(\sum_{i=1}^{N}\exp\left(\gamma_2 R\left(c_i, S_{local'_i}\right)\right)\right)^{\frac{1}{\gamma_2}}, \tag{15}$$

where γ_1 and γ_2 are hyperparameters. The global attribute feature S_{global} and global image feature i_{global} are employed to calculate another matching degree between the generated image and input attribute in terms of global evaluation:

$$R^{global}(Q,D) = \frac{i_{global}^T S_{global}}{\|i_{global}\|\|S_{global}\|}. \tag{16}$$

This two matching degrees can be used to evaluate the quality of the generated image and reflect whether it matches the input attributes.

E. Objective function

The objective function of the whole network is defined as follows:

$$\mathscr{L} = \mathscr{L}_G + \lambda\mathscr{L}_{SCM}, \quad \mathscr{L}_G = \sum_{i=0}^{2}\mathscr{L}_{G_i}, \tag{17}$$

where $\mathscr{L}_G$ denotes the objective functions from three stacked image generation modules and $\mathscr{L}_{SCM}$ denotes the objective function in SCM module.

The adversarial loss function for G_i is defined as follows:

$$\mathscr{L}_{G_i} = -\frac{1}{2}\mathbb{E}_{x_i\sim p_{G_i}}\left[\log\left(D_i\left(x_i\right)\right)\right] - \frac{1}{2}\mathbb{E}_{x_i\sim p_{G_i}}\left[\log\left(D_i\left(x_i, S_{global}\right)\right)\right]. \tag{18}$$

The adversarial loss function for D_i is defined as follows:

$$\begin{aligned}\mathscr{L}_{D_i} = &-\frac{1}{2}\mathbb{E}_{x_i^{gt}\sim p_{data_i}}\left[\log D_i\left(x_i^{gt}\right)\right] - \frac{1}{2}\mathbb{E}_{x_i\sim p_{G_i}}\left[\log\left(1-D_i\left(x_i\right)\right)\right] \\ &-\frac{1}{2}\mathbb{E}_{x_i^{gt}\sim p_{data_i}}\left[\log D_i\left(x_i^{gt}, S_{global}\right)\right] \\ &-\frac{1}{2}\mathbb{E}_{x_i\sim p_{G_i}}\left[\log\left(1-D_i\left(x_i, S_{global}\right)\right)\right],\end{aligned} \tag{19}$$

where x_i^{gt} is the ground truth image in the corresponding generator.

For the objective function in SCM module, we first use the local matching degree $R^{local}(Q,D)$ to calculate the matching degree between the generated image and input attribute for each input attribute in a minibatch of training data:

$$P\left(D_i|Q_i\right) = \frac{\exp\left(\gamma_3 R^{local}\left(Q_i, D_i\right)\right)}{\sum_{j=1}^{M}\exp\left(\gamma_3 R^{local}\left(Q_i, D_j\right)\right)}, \tag{20}$$

where γ_3 is a hyperparameter and M is the number of samples in a minibatch of training data. And the corresponding objective function can be defined as follows:

$$\mathscr{L}_1^{local} = -\sum_{i=1}^{M}\log P\left(D_i|Q_i\right). \tag{21}$$

Similarly, we can defined the matching degree between the generated image and input attribute for each generated image in a minibatch of training data, and calculate the corresponding objective function:

$$P\left(Q_i|D_i\right) = \frac{\exp\left(\gamma_3 R^{local}\left(Q_i, D_i\right)\right)}{\sum_{j=1}^{M}\exp\left(\gamma_3 R^{local}\left(Q_j, D_i\right)\right)}, \tag{22}$$

$$\mathscr{L}_2^{local} = -\sum_{i=1}^{M}\log P\left(Q_i|D_i\right). \tag{23}$$

Similarly by using the global matching loss $R^{global}(Q,D)$, we can correspondingly achieve $\mathscr{L}_1^{global}$ and $\mathscr{L}_2^{global}$. Finally, the objective function of SCM is defined as follows:

$$\mathscr{L}_{SCM} = \mathscr{L}_1^{local} + \mathscr{L}_2^{local} + \mathscr{L}_1^{global} + \mathscr{L}_2^{global}. \tag{24}$$

IV. Experiment

In this section, we will demonstrate the superiority of our AFGAN through both qualitative and quantitative evaluations. We first introduce the experiments setting and the implementation details of our method.

Then we present the ablation study of our method to investigate the effectiveness of each module proposed in AFGAN. Furthermore, we compare AFGAN with the state-of-the-art method AttnGAN [1] in terms of qualitative evaluations. Finally we employ common metrics (e.g. BRISQUE, IS, FID and MS-SSIM) and an extra trained facial attribute predictor to do quantitative evaluations.

A. Experimental Settings

We use CelebA [22] dataset to evaluate our proposed AFGAN. CelebA dataset consists of 202599 face images, and each has 40 kinds of facial attributes. After performing face detection on all faces, 180694 faces are treated as the training set and 19761 faces are used for testing. Since some attributes can not be recognized from a tight face region (such as necklaces) or some attributes are difficult to measure (such as attractive), we select 18 attributes from them for our experiments. The selected 18 attributes are listed in Table I.

We compare our method with the state-of-the-art method AttnGAN[1] in terms of both qualitative and quantitative evaluations. Since the AttnGAN is a text-to-image model, and the input of our work is attribute vector instead of text, we change the onehot text vector into onehot attribute vector as the input of model.

TABLE I
18 ATTRIBUTES SELECTED FROM CELEBA DATASET.

No.	Attribute	No.	Attribute
0	5_o_Clock_Shadow	9	Eyeglasses
1	Arched_Eyebrows	10	Gray_Hair
2	Bags_Under_Eyes	11	Male
3	Bald	12	Mouth_Slightly_Open
4	Bangs	13	Narrow_Eyes
5	Black_Hair	14	No_Beard
6	Blond_Hair	15	Pale_Skin
7	Brown_Hair	16	Pointy_Nose
8	Bushy_Eyebrows	17	Smiling

B. Implementation details

Our proposed AFGAN consists of three modules: AEM, SIGM and SCM.

We first pretrain the image encoder (as shown in Figure 2) in SCM. The SCM module is used to calculate the similarity between the input attribute features and the generated image features. Specifically, the input attributes features consist of the global and local features obtained from the AEM module, while the generated image features are obtained by AlexNet, which is a pretrained attribute prediction model. The local features of the generated image are the features before the third pooling layer in AlexNet, and the global features of the generated image come from the last full connection layer. After getting the global and local features of attributes and images respectively, the loss function is calculated according to formula (24).

The AEM module is composed of global and local parts. The global part is to extract the feature of the input attribute vector through a full connection layer. The local part is mapped by two embedding tables first, as shown in Figure 3. Then a self-attention layer is carried out by using the formula of $softmax(AA^T)A$, where A is the feature generated from the previous two-path embedding layer, so that the relationship between different attributes can be integrated.

Traditional convolution network is employed as discriminators in the first two stages in SIGM to discriminate the authenticity of generated facial images. And we use PatchGAN [23] as discriminator in the third stage since it can pay more attention to details of the generated images. We adopt WGAN_GP [24] loss as the GAN loss.

We use Adam optimizer [25] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for training the network. We set the attribute feature dimension C as 256 and the dimension of noise vector z as 100. We adjust the update frequency of generator and discriminator during the training of GAN, i.e., updating once discriminator every four times of generator. We train AFGAN model with 30 epochs on 3 GTX 1080Ti GPUs for about 5 days.

C. Ablation Study

To demonstrate the superiority of our AFGAN, we do ablation study to investigate the effectiveness of each module proposed in AFGAN.

1) SCM Module: The SCM module uses the pretrained encoder to constrain the distance between the features of attribute vector and the generated image. After the training of the SCM module, α in SCM represents the correlation between the image and each of input attributes. For each facial image, the attention map for the response of each attribute is shown in Figure 4, where the red label means the value of the attribute is 1 and the black means the value of the attribute is 0. The white region of the attention map represents the response area corresponding to the attribute. We can see that the response areas of most attributes are reasonable, such as hair (black hair, brown hair), nose, etc. In addition, we can see that the response areas of the five-facial points attributes like nose, month and eyes are smaller, while that of the attributes such as smile, mouth opening and hair are larger. This phenomenon is also consistent with our common sense. The facial regions involved in the five-facial points only concentrate on the five-facial points' locations themselves, but there are no fixed areas for the attributes such as hair, smile and so on. Moreover paying attention to the attributes of male and bangs, we can see that when the value of the attribute is 0 (i.e., the label in the figure is black), the attention map does not all turn black, but still responds to certain areas. It indicates that whether the value of the attribute is 0 or 1, the generated image always has corresponding areas to well reflect its meaning, which demonstrates the effectiveness of the proposed two-path embedding layer.

We can also compare the generated images in SIGM module on condition of with and without SCM module in Figure 1. The second row represents the experiment results without the constraints of SCM module in the objective function, and third row represents the results of the complete AFGAN model proposed by us. The input attribute label is shown in the top color block containing the attribute number and the corresponding content. The ground truth images corresponding to the attribute labels are shown in the last row. It can be seen that the quality of the image generated by our proposed AFGAN model with SCM module is clearer and more realistic than that of without SCM module. The second row shows the generated images with two-path embedding layer and self-attention layer, and there are some distortion phenomena due to the lack of SCM module. By adding the constraints of SCM module, the generated images in third row make progresses in terms of diversity and authenticity, and reflect the details of facial texture well.

2) The generated image of three stages in SIGM module: Figure 5 shows the images generated by three stages of SIGM module separately. The images smaller than 256×256 pixels are enlarged to 256×256 pixels by bilinear interpolation. The left-most of each group is the ground truth image with the corresponding label, and the three images on the right are 64×64, 128×128 and 256×256 pixels respectively. We can see that the 64×64 pixel image generated in the first stage mainly reflects the overall attribute of the face, such as gender, hairstyle. Then some fine-grained texture information is gradually reflected in the following two stages, such as wrinkles of the face, texture of the hair, lightness of the face and so on. Moreover, the faces generated in the three stages are consistent. The faces generated in latter stage do not change

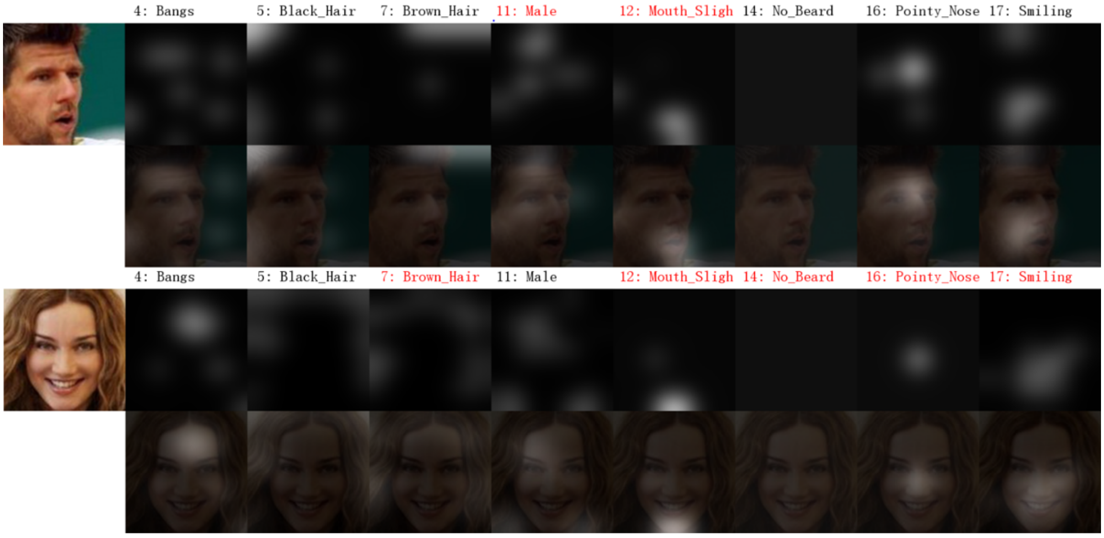

Fig. 4. The attention maps in SCM module. The red label means the value of the attribute is 1 and the black means the value of the attribute is 0. The white region of the attention map represents the response area corresponding to the attribute.

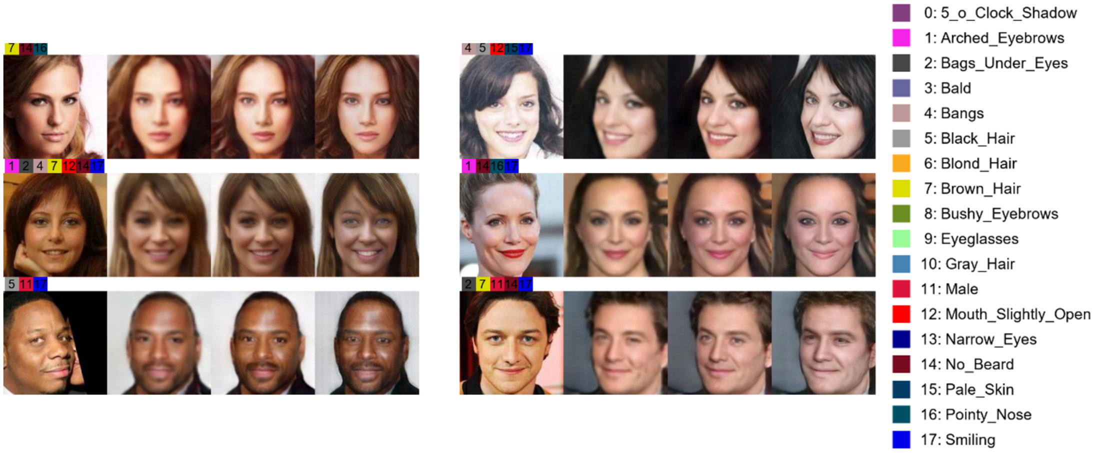

Fig. 5. The generated face images of three stages in SIGM module. In each set of four images, the left-most of each group is the ground truth image of the corresponding label, and the three images on the right are 64×64, 128×128 and 256×256 pixels respectively. The images smaller than 256×256 pixels are enlarged to 256×256 pixels by bilinear interpolation. The input attribute label is shown in the top color block and number, and the corresponding attribute content is listed on the right side.

the identity information in the previous stage, but only improve the details of the face. In this way, the facial image can be generated step by step in three stages according to the input attribute label, which illustrates that SIGM module is effective.

D. *Qualitative evaluations*

We compare the images generated by AttnGAN [1] and our proposed AFGAN in Figure 1. The images generated by our method are randomly selected. The first row represents the results of AttnGAN model. The second row represents the results of AFGAN model proposed by us. The ground truth images corresponding to the attribute labels are shown in the last row. The generated face images by AttnGAN are blurred and have the phenomenon of mode collapse [2]. On the contrary, the images generated by AFGAN greatly alleviate the phenomenon of mode collapse and distortion.

E. *Quantitative evaluations*

Moreover, we employ BRISQUE [26], IS [27], FID [28] and MS-SSIM [29] to further evaluate our model quantitatively. BRISQUE calculates the no-reference image quality score for images using the Blind/Referenceless Image Spatial Quality Evaluator. A smaller score indicates better perceptual quality. IS and FID are typical image evaluation metrics of the GAN model, which respectively focus on the diversity of generated images and the feature distance between the generated images and the real images. MS-SSIM means multi-scale structural similarity approach for image quality assessment, which provides more flexibility than single-scale approach in incorporating the variations of image resolution and viewing conditions. For AttnGAN, we use attribute vector instead of text as input. The results are shown in Table II.

TABLE II
MORE QUANTITATIVE ANALYSIS BY METRIC OF BRISQUE, IS, FID AND MS-SSIM.

	BRISQUE ↓	IS ↑	FID ↓	MS-SSIM ↓
AttnGAN [1]	62.843	5.124	40.254	0.398
Wang et al. [11]	—	2.2	43.8	—
AFGAN(ours)	35.979	5.853	36.607	0.347

For the metric of BRISQUE, AFGAN achieves 35.979, which is much lower than the result of AttnGAN, i.e., 62.843. It means that the quality of generated images from AFGAN is much better than AttnGAN. For the metric of MS-SSIM, AFGAN achieves 0.347 which is much lower than 0.398 obtained by AttnGAN. The lower value of MS-SSIM means the generated images are more diverse. For the metric of IS, AFGAN achieves a better result of 5.853 compared to 5.124 achieved by AttnGAN and 2.2 by Wang et al., which demonstrates that the images generated by AFGAN have better image quality again. For the metric of FID, AFGAN achieves 36.607 while AttnGAN and Wang et al. achieve 40.254 and 43.8 respectively, which means that the images generated by AFGAN show better authenticity and are closer to the real images. All the results show the superiority of our AFGAN over AttnGAN [1] and Wang et al. [11] for generating high quality faces.

Besides, we use the CelebA dataset [22] to train an attribute prediction model. Then we use the generated images by our model to test the classification accuracy compared to the corresponding input facial attributes. The results of AttnGAN, AFGAN without AEM, AFGAN without SCM and AFGAN are shown in Table III. As seen, AFGAN achieves the highest classification accuracy, which means that the images generated by AFGAN can better reflect the input attributes.

TABLE III
THE CLASSIFICATION ACCURACY OF GENERATED FACIAL IMAGES BY DIFFERENT SETTING.

Setting	Classification accuracy
AttnGAN	0.902
AFGAN w/o AEM	0.924
AFGAN w/o SCM	0.940
AFGAN(ours)	0.955

V. Conclusion

In this work, we propose a novel attributes aware face image generation method with generative adversarial networks called AFGAN to solve the problem of generating facial image with specific attributes. AFGAN is composed of three modules, i.e., AEM, SIGM and SCM. AEM uses a two-path embedding layer and self-attention layer to convert binary attribute vector to rich attribute features. SIGM conducts three stacked generators to generate 64×64, 128×128 and 256×256 resolution faces respectively. And SCM propose an image-attribute matching loss to enhance the correlation between the generated images and input attributes. Both qualitative and quantitative evaluations on CelebA dataset show the effectiveness of AFGAN.

VI. Acknowledgment

This work is partially supported by Natural Science Foundation of China (Nos. 61806188, 61976219), and the state key development program in 13th Five-Year under Grant (No. Y808401).

References

[1] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *CVPR*, 2018, pp. 1316–1324.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[3] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, 2014.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *arXiv:1812.04948*, 2018.

[5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *arXiv:1912.04958*, 2019.

[6] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis," in *CVPR*, 2019, pp. 5802–5810.

[7] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, and J. Gao, "Object-driven text-to-image synthesis via adversarial training," in *CVPR*, 2019, pp. 12 174–12 182.

[8] T. Hinz, S. Heinrich, and S. Wermter, "Semantic object accuracy for generative text-to-image synthesis," *arXiv:1910.13321*, 2019.

[9] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional cyclegan," in *ECCV*, 2018, pp. 282–297.

[10] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *ECCV*. Springer, 2016, pp. 776–791.

[11] Y. Wang, A. Dantcheva, and F. Bremond, "From attribute-labels to faces: face generation using a conditional generative adversarial network," in *ECCV*, 2018.

[12] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv:1805.08318*, 2018.

[13] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv:1710.10196*, 2017.

[14] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017, pp. 1501–1510.

[15] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," *arXiv:1903.05854*, 2019.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.

[17] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes," *IEEE TPAMI*, 2019.

[18] M. Li, Y. Sun, Z. Zhang, H. Xie, and J. Yu, "Deep learning face hallucination via attributes transfer and enhancement," in *IEEE ICME*. IEEE, 2019, pp. 604–609.

[19] X. Di and V. M. Patel, "Face synthesis from visual attributes via sketch using conditional vaes and gans," 2017.

[20] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017, pp. 5907–5915.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[22] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[24] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *NIPS*, 2017, pp. 5767–5777.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[26] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE TIP*, vol. 21, no. 12, pp. 4695–4708, 2012.

[27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NIPS*, 2016, pp. 2234–2242.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NIPS*, 2017, pp. 6626–6637.

[29] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.