

# AttGAN: Facial Attribute Editing by Only Changing What You Want

Zhenliang He, Wangmeng Zuo, *Senior Member, IEEE*, Meina Kan, *Member, IEEE*,  
Shiguang Shan, *Senior Member, IEEE*, and Xilin Chen, *Fellow, IEEE*

**Abstract**—Facial attribute editing aims to manipulate single or multiple attributes of a face image, i.e., to generate a new face with desired attributes while preserving other details. Recently, generative adversarial net (GAN) and encoder-decoder architecture are usually incorporated to handle this task with promising results. Based on the encoder-decoder architecture, facial attribute editing is achieved by decoding the latent representation of the given face conditioned on the desired attributes. Some existing methods attempt to establish an attribute-independent latent representation for further attribute editing. However, such attribute-independent constraint on the latent representation is excessive because it restricts the capacity of the latent representation and may result in information loss, leading to over-smooth and distorted generation. Instead of imposing constraints on the latent representation, in this work we apply an *attribute classification constraint* to the generated image to just guarantee the correct change of desired attributes, i.e., to “change what you want”. Meanwhile, the *reconstruction learning* is introduced to preserve attribute-excluding details, in other words, to “only change what you want”. Besides, the *adversarial learning* is employed for visually realistic editing. These three components cooperate with each other forming an effective framework for high quality facial attribute editing, referred as AttGAN. Furthermore, our method is also directly applicable for *attribute intensity control* and can be naturally extended for *attribute style manipulation*. Experiments on CelebA dataset show that our method outperforms the state-of-the-arts on realistic attribute editing with facial details well preserved.

**Index Terms**—facial attribute editing, attribute intensity control, attribute style manipulation, adversarial learning

## I. INTRODUCTION

**T**HIS work investigates the facial attribute editing task, which aims to edit a face image by manipulating single or multiple attributes of interest (e.g., hair color, expression, mustache and age). For conventional face recognition [1], [2] and facial attribute prediction [3], [4] tasks, significant advances have been made along with the development of deep convolutional neural networks (CNNs) and large scale labeled datasets. However, it is difficult or even impossible to collect labeled images of a same person with varying attributes, thus supervised learning is generally inapplicable for facial attribute editing. Therefore, researchers turn to generative models such as variational autoencoder (VAE) [5] and generative adversarial network (GAN) [6], and make considerable progress on facial attribute editing [7]–[16].

Some existing methods [9]–[12] use different editing models for different attributes, therefore one has to train numerous models for handling various attribute editing subtasks, which is difficult for real deployment. For this problem, the encoder-decoder architecture [7], [8], [13]–[15] seems to be an effective

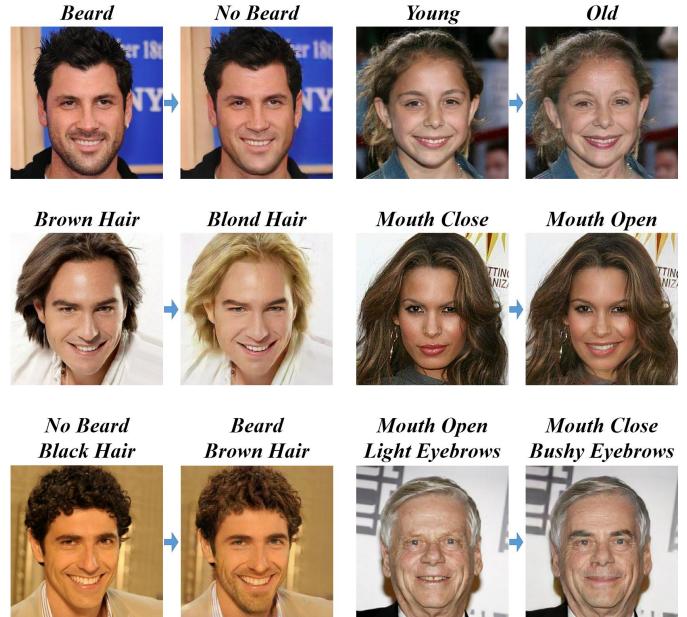


Fig. 1. Facial attribute editing results from our AttGAN. Zoom in for better resolution.

solution for using a *single* model for *multiple* attribute manipulation. Therefore, we also focus on the encoder-decoder architecture and develop an effective method for high quality facial attribute editing.

With the encoder-decoder architecture, facial attribute editing is achieved by decoding the latent representation from the encoder conditioned on the expected attributes. Based on such framework, the key issue of facial attribute editing is **how to model the relation between the attributes and the face latent representation**. For this issue, VAE/GAN [7] represents each attribute as a vector, which is defined as the difference between the mean latent representations of the faces with and without this attribute. Then, by adding a single or multiple attribute vectors to a face latent representation, the decoded face image from the modified representation is expected to own those attributes. However, such attribute vector contains highly correlated attributes, thus inevitably leading to unexpected changes of other attributes, e.g., adding blond hair always makes a male become a female because most blond hair objects are female in the training set. In IcGAN [8], the latent representation is sampled from a normal distribution independent of the attributes. In Fader Networks [13], an adversarial process is introduced to force the latent representation

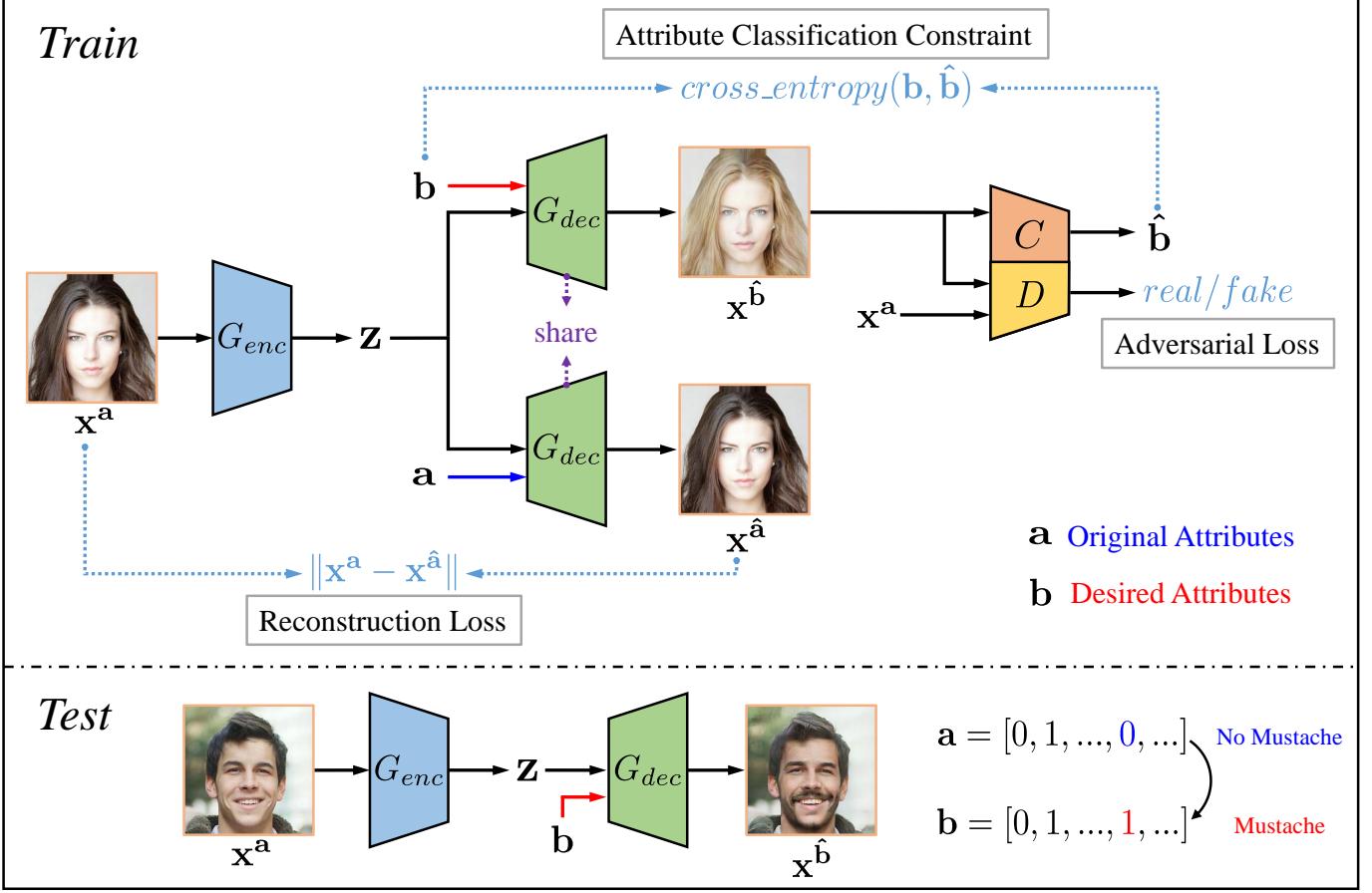


Fig. 2. Overview of our AttGAN, which contains three main components at training: the attribute classification constraint, the reconstruction learning and the adversarial learning. The attribute classification constraint guarantees the correct attribute manipulation on the generated image. The reconstruction learning aims at preserving the attribute-excluding details. The adversarial learning is employed for visually realistic generation.

to be invariant to the attributes. However, the attributes portray the characteristics of a face image, which implies the relation between the attributes and the face latent representation is highly complex and closely dependent. Therefore, simply imposing the attribute-independent constraint on the latent representation not only restricts its representation ability but also may result in information loss, which is harmful to the attribute editing.

With the above limitation analysis of existing methods in mind, we argue that the invariance of the latent representation to the attributes is excessive, and what we need is just the correct editing of attributes. To this end, instead of imposing the attribute-independence constraint on the latent representation [8], [13], we apply an attribute classification constraint to the generated image, just requiring the correct attribute manipulations, i.e., to “change what you want”. Therefore in comparison with IcGAN [8] and Fader Networks [13], the latent representation in our method is constraint free, which guarantees its representation ability and flexibility for further attribute editing. Besides, we introduce the reconstruction learning for the preservation of the attribute-excluding details<sup>1</sup>, i.e., we aim to “only change” the expected attributes while keeping the other details unchanged. Moreover, the adversarial learning is employed for visually realistic editing.

<sup>1</sup>attribute-excluding details mean the other details of a face image except for the expected attributes, such as face identity, illumination and background.

Our method, referred as AttGAN, can generate visually more pleasing results with fine facial details (see Fig. 1) in comparison with the state-of-the-arts. Moreover, our AttGAN is directly applicable for attribute intensity control and can be naturally extended for attribute style manipulation. To sum up, the contribution of this work lies in three folds:

- Properly considering the relation between the attributes and the face latent representation under the principle of just satisfying the correct editing objective. Our AttGAN removes the strict attribute-independent constraint from the latent representation, and just applies the attribute classification constraint to the generated image to guarantee the correct change of the attributes.
- Incorporating the attribute classification constraint, the reconstruction learning and the adversarial learning into a unified framework for high quality facial attribute editing, i.e., the attributes are correctly edited, the attribute-excluding details are well preserved and the whole image is visually realistic.
- Promising results of multiple facial attribute editing using a single model. AttGAN outperforms the state-of-the-arts with better perceptual quality for facial attribute editing. Moreover, our method is directly applicable for attribute intensity control and can be naturally extended for attribute style manipulation.

## II. RELATED WORK

### A. Facial Attribute Editing

There are two types of methods for facial attribute editing, the optimization based ones [17], [18] and the learning based ones [7]–[14], [16]. Optimization based methods include CNAI [17] and DFI [18]. To change a given face to the target face with the expected attributes, CNAI [17] defines an attribute loss as the CNN feature difference between the given face and a set of faces with the expected attributes, and then minimizes this loss with respect to the given face. Based on the assumption that CNN linearizes the manifold of the natural images into an Euclidean feature subspace [19], DFI [18] first linearly moves the deep feature of the input face along the direction vector between the faces with and without the expected attributes. Then the facial attribute editing is achieved by optimizing the input face to match its deep feature with the moved feature. Optimization based methods need to conduct several or even many optimization iterations for each testing image, which are usually time-consuming and unfriendly for real world applications.

Learning based methods are more popular. Li et al. [9] present to train a deep identity-aware attribute transfer model to add/remove an attribute to/from a face image by employing an adversarial attribute loss and a deep identity feature loss. Shen and Liu [10] adopt the dual residual learning strategy to simultaneously train two networks for respectively adding and removing a specific attribute. GeneGAN [12] swaps a specific attribute between two given images by recombining the information of their latent representation. These methods [9]–[12], however, train different models for different attributes (or attribute combinations), leading to large number of models which are also unfriendly for real world applications.

Several learning based methods have been proposed for multiple facial attribute editing with one model. In VAE/GAN [7], GAN [6] and VAE [5] are combined to learn a latent representation and a decoder. Then the attribute editing is achieved by modifying the latent representation to own the information of expected attributes and then decoding it. IcGAN [8] separately trains a cGAN [20] and an encoder, requiring that the latent representation is sampled from a uniform distribution and therefore independent of the attributes. Then the attribute editing is performed by first encoding an image into the latent representation and then decoding the representation conditioned on the given attributes. Fader Networks [13] employs an adversarial process on the latent representation of an autoencoder to learn the attribute-invariant representation. Then, the decoder takes such representation and arbitrary attribute vector as input to generate the edited result. However, the attribute-independent constraint on the latent representation in IcGAN and Fader Networks is excessive, because it harms the representation ability and may result in information loss, leading to unexpected distortion on the generated images (e.g., over smoothing). Kim et al. [14] define different blocks of the latent code as the representations of different attributes, and swap several latent code blocks between two given images to achieve multiple attribute swapping. DNA-GAN [15] also swap attribute relevant latent

blocks between a given pair of images to make “crossbreed” images. Both Kim et al. [14] and DNA-GAN [15] can be viewed as extensions of GeneGAN [12] for multiple attributes. StarGAN [16] trains a conditional attribute transfer network via attribute classification loss and cycle consistency loss. StarGAN and our AttGAN are concurrently and independently proposed<sup>2</sup> and share some similar objective functions. Main differences between StarGAN and AttGAN are in two folds: 1) StarGAN uses cycle consistency loss and AttGAN does not include cyclic process or cycle consistency loss, 2) StarGAN trains a conditional attribute transfer network and does not involve any latent representation while AttGAN uses an encoder-decoder architecture and models the relation between the latent representation and the attributes.

Image translation task is closely related to facial attribute editing and some image translation methods are also directly applicable for facial attribute editing. CycleGAN [21] trains two bidirectional transfer models between two image domains by employing the cycle consistency loss and two domain specific adversarial learning processes. UNIT [11] learns to encode the images of two different domains into a common latent space, and then decode the latent representation to the expected domain via the domain specific decoder. Separating face images with and without the expected attributes into two different domain, one can directly use these methods for facial attribute editing. However, inability of handling multiple attributes with single model is also the limitation of these domain translation methods.

Our AttGAN is a learning based method for single or multiple facial attribute editing, which is mostly motivated by the encoder-decoder based methods VAE/GAN [7], IcGAN [8] and Fader Networks [13]. We mainly focus on the disadvantages of these three methods on modeling the relation between the latent representation and the attributes, and propose a novel method to solve such problem.

### B. Generative Adversarial Networks

Denote by  $p_{data}(\mathbf{x})$  the distribution of the real image  $\mathbf{x}$ , and  $p_{\mathbf{z}}(\mathbf{z})$  the distribution of the input. Generative adversarial net (GAN) [6] is a special generative model to learn a generator  $G(\mathbf{z})$  to capture the distribution  $p_{data}$  via an adversarial process. Specifically, a discriminator  $D$  is introduced to distinguish the generated images from the real ones, while the generator  $G(\mathbf{z})$  is updated to confuse the discriminator. The adversarial process is formulated as a minimax game as

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{data}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} [\log(1 - D(G(\mathbf{z})))] \quad (1)$$

Theoretically, when the adversarial process reaches the Nash equilibrium, the minimax game attains its global optimum  $p_{G(\mathbf{z})} = p_{data}$  [6].

GAN is notorious for its unstable training and mode collapse. DCGAN [22] uses CNN and batch normalization [23] for stable training. Subsequently, to avoid mode collapse and further enhance the training stability, WGAN [24] minimizes

<sup>2</sup>StarGAN first appears on 2017.11.24 - <http://arxiv.org/abs/1711.09020>, and our AttGAN first appears on 2017.11.29 - <http://arxiv.org/abs/1711.10678>.

the Wasserstein-1 distance between the generated distribution and the real distribution as

$$\min_G \max_{\|D\|_L \leq 1} \mathbb{E}_{\mathbf{x} \sim p_{data}} [D(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} [D(G(\mathbf{z}))], \quad (2)$$

where  $D$  is constrained to be the 1-Lipschitz function implemented by weight clipping. Furthermore, WGAN-GP [25] improves WGAN on the implementation of Lipschitz constraint by imposing a gradient penalty on the discriminator instead of weight clipping. In this work, we adopt WGAN-GP for the adversarial learning.

Several works have been developed for the conditional generation with given attributes or class labels [20], [26]–[28]. Employing an auxiliary classifier or regressor, both AC-GAN [27] and InfoGAN [28] learn the conditional generation by mapping the generated images back to the conditional signals. Inspired, in this work, we also map the edited face images back to the given attributes forming the attribute classification constraint. Different from AC-GAN [27], the generated images do not participate in the training of the auxiliary classifier.

### III. ATTRIBUTE GAN (ATTGAN)

This section introduces the AttGAN approach for the editing of binary facial attributes<sup>3</sup>. As shown in Fig. 2, our AttGAN is comprised of two basic subnetworks, i.e., an encoder  $G_{enc}$  and a decoder  $G_{dec}$ , together with an attribute classifier  $C$  and a discriminator  $D$ . In the following, we describe the design principles of AttGAN and introduce the objectives for training these components. Then we present an extension of AttGAN for attribute style manipulation.

#### A. Testing Formulation

Given a face image  $\mathbf{x}^{\mathbf{a}}$  with  $n$  binary attributes  $\mathbf{a} = [a_1, \dots, a_n]$ , the encoder  $G_{enc}$  is used to encode  $\mathbf{x}^{\mathbf{a}}$  into the latent representation, denoted as

$$\mathbf{z} = G_{enc}(\mathbf{x}^{\mathbf{a}}). \quad (3)$$

Then the process of editing the attributes of  $\mathbf{x}^{\mathbf{a}}$  to another attributes  $\mathbf{b} = [b_1, \dots, b_n]$  is achieved by decoding  $\mathbf{z}$  conditioned on  $\mathbf{b}$ , i.e.,

$$\mathbf{x}^{\hat{\mathbf{b}}} = G_{dec}(\mathbf{z}, \mathbf{b}), \quad (4)$$

where  $\mathbf{x}^{\hat{\mathbf{b}}}$  is the edited image expected to own the attribute  $\mathbf{b}$ . Thus the whole editing process is formulated as

$$\mathbf{x}^{\hat{\mathbf{b}}} = G_{dec}(G_{enc}(\mathbf{x}^{\mathbf{a}}), \mathbf{b}). \quad (5)$$

#### B. Training Formulation

It can be seen from Eq. (5) that the attribute editing problem can be formally defined as the learning of the encoder  $G_{enc}$  and decoder  $G_{dec}$ . This learning problem is unsupervised, because the ground truth of the editing, i.e.  $\mathbf{x}^{\mathbf{b}}$ , is unavailable.

On one hand, the editing on the given face image  $\mathbf{x}^{\mathbf{a}}$  is expected to produce a realistic image with attributes  $\mathbf{b}$ . For

<sup>3</sup>each attribute is represented by 1/0 for with/without it and all attributes are represented by a 1/0 sequence.

this purpose, an **attribute classifier** is used to constrain the generated image  $\mathbf{x}^{\hat{\mathbf{b}}}$  to correctly own the desired attributes, i.e., the attribute prediction of  $\mathbf{x}^{\hat{\mathbf{b}}}$  should be  $\mathbf{b}$ . Meanwhile, the **adversarial learning** is employed on  $\mathbf{x}^{\hat{\mathbf{b}}}$  to ensure its visual reality.

On the other hand, an eligible attribute editing should only change those desired attributes, while keeping the other details unchanged. To this end, the **reconstruction learning** is introduced to 1) make the latent representation  $\mathbf{z}$  conserve enough information for the later recovery of the attribute-excluding details, 2) enable the decoder  $G_{dec}$  to restore the attribute-excluding details from  $\mathbf{z}$ . Specifically, for the given  $\mathbf{x}^{\mathbf{a}}$ , the generated image conditioned on its own attributes  $\mathbf{a}$ , i.e.,

$$\mathbf{x}^{\hat{\mathbf{a}}} = G_{dec}(\mathbf{z}, \mathbf{a}) \quad (6)$$

should approximate  $\mathbf{x}^{\mathbf{a}}$  itself, i.e.,  $\mathbf{x}^{\hat{\mathbf{a}}} \rightarrow \mathbf{x}^{\mathbf{a}}$ .

In summary, the relation between the attributes  $\mathbf{a}/\mathbf{b}$  and the latent representation  $\mathbf{z}$  is implicitly modeled in two aspects: 1) the interaction between  $\mathbf{z}$  and  $\mathbf{b}$  in the decoder should produce an realistic image  $\mathbf{x}^{\hat{\mathbf{b}}}$  with correct attributes, and 2) the interaction between  $\mathbf{z}$  and  $\mathbf{a}$  in the decoder should produce an image  $\mathbf{x}^{\hat{\mathbf{a}}}$  approximating the input  $\mathbf{x}^{\mathbf{a}}$  itself.

**Attribute Classification Constraint.** As mentioned above, it is required that the generated image  $\mathbf{x}^{\hat{\mathbf{b}}}$  should correctly own the new attributes  $\mathbf{b}$ . Therefore, we employ an attribute classifier  $C$  to constrain the generated image  $\mathbf{x}^{\hat{\mathbf{b}}}$  to own the desired attributes, i.e.,  $C(\mathbf{x}^{\hat{\mathbf{b}}}) \rightarrow \mathbf{b}$ , formulated as follows,

$$\min_{G_{enc}, G_{dec}} \mathcal{L}_{cls_g} = \mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim p_{data}, \mathbf{b} \sim p_{attr}} [\ell_g(\mathbf{x}^{\mathbf{a}}, \mathbf{b})], \quad (7)$$

$$\ell_g(\mathbf{x}^{\mathbf{a}}, \mathbf{b}) = \sum_{i=1}^n -b_i \log C_i(\mathbf{x}^{\hat{\mathbf{b}}}) - (1 - b_i) \log(1 - C_i(\mathbf{x}^{\hat{\mathbf{b}}})), \quad (8)$$

where  $p_{data}$  and  $p_{attr}$  indicate the distribution of real images and the distribution of attributes,  $C_i(\mathbf{x}^{\hat{\mathbf{b}}})$  indicates the prediction of the  $i^{th}$  attribute, and  $\ell_g(\mathbf{x}^{\mathbf{a}}, \mathbf{b})$  is the summation of binary cross entropy losses of all attributes.

The attribute classifier  $C$  is trained on the input images with their original attributes, by the following objective,

$$\min_C \mathcal{L}_{cls_c} = \mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim p_{data}} [\ell_r(\mathbf{x}^{\mathbf{a}}, \mathbf{a})], \quad (9)$$

$$\ell_r(\mathbf{x}^{\mathbf{a}}, \mathbf{a}) = \sum_{i=1}^n -a_i \log C_i(\mathbf{x}^{\mathbf{a}}) - (1 - a_i) \log(1 - C_i(\mathbf{x}^{\mathbf{a}})). \quad (10)$$

**Reconstruction Loss.** Furthermore, the reconstruction learning aims for satisfactory preservation of attribute-excluding details. To this end, the decoder should learn to reconstruct the input image  $\mathbf{x}^{\mathbf{a}}$  by decoding the latent representation  $\mathbf{z}$  conditioned on the original attributes  $\mathbf{a}$ . The learning objective is formulated as

$$\min_{G_{enc}, G_{dec}} \mathcal{L}_{rec} = \mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim p_{data}} [\|\mathbf{x}^{\mathbf{a}} - \mathbf{x}^{\hat{\mathbf{a}}}\|_1], \quad (11)$$

where we use the  $\ell_1$  loss rather than  $\ell_2$  loss to suppress the blurriness.

**Adversarial Loss.** The adversarial learning between the generator (including the encoder and decoder) and discriminator is introduced to make the generated image  $\mathbf{x}^{\hat{\mathbf{b}}}$  visually realistic. Following WGAN [24], the adversarial losses for the discriminator and generator are formulated as below,

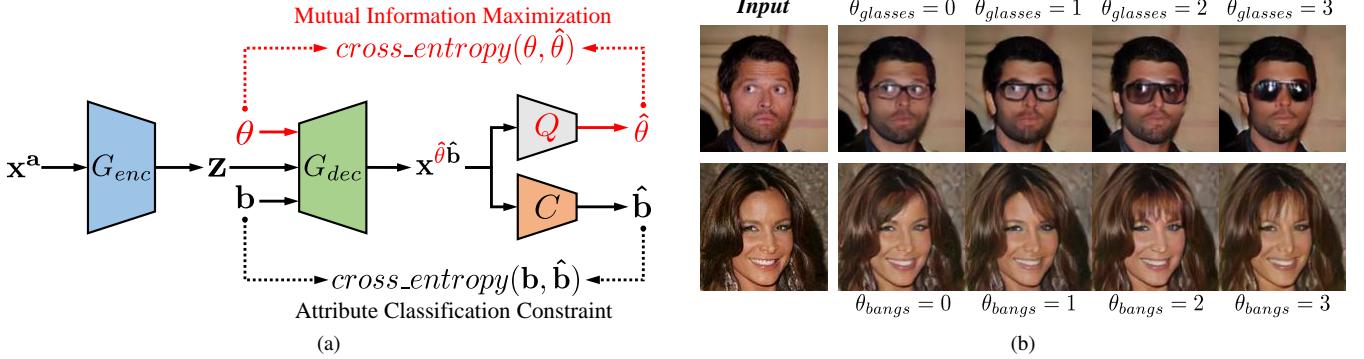


Fig. 3. Illustration of AttGAN extension for attribute style manipulation. (a) shows the extended framework based on the original AttGAN.  $\theta$  denotes the style controllers and  $Q$  denotes the style predictor. (b) shows the visual effect of changing attribute style by varying  $\theta$ .

$$\min_{\|D\|_L \leq 1} \mathcal{L}_{adv_d} = -\mathbb{E}_{x^a \sim p_{data}} D(x^a) + \mathbb{E}_{x^a \sim p_{data}, b \sim p_{attr}} D(x^{\hat{b}}), \quad (12)$$

$$\min_{G_{enc}, G_{dec}} \mathcal{L}_{adv_g} = -\mathbb{E}_{x^a \sim p_{data}, b \sim p_{attr}} [D(x^{\hat{b}})], \quad (13)$$

where  $D$  is the discriminator described in Eq. (2). The adversarial losses are optimized via WGAN-GP [25].

**Overall Objective.** By combining the attribute classification constraint, the reconstruction loss and the adversarial loss, an unified attribute GAN (AttGAN) is obtained, which can edit the desired attributes with the attribute-excluding details well preserved. Overall, the objective for the encoder and decoder is formulated as below,

$$\min_{G_{enc}, G_{dec}} \mathcal{L}_{enc, dec} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{cls_g} + \mathcal{L}_{adv_g}, \quad (14)$$

and the objective for the discriminator and the attribute classifier is formulated as below,

$$\min_{D, C} \mathcal{L}_{dis, cls} = \lambda_3 \mathcal{L}_{cls_c} + \mathcal{L}_{adv_d}, \quad (15)$$

where the discriminator and the attribute classifier share most layers,  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  are the hyperparameters for balancing the losses.

### C. Why are attribute-excluding details preserved?

The above AttGAN design can be viewed as a multi-task leaning of attribute editing task with classification loss and face reconstruction task with reconstruction loss, which share the entire encoder-decoder network. However, AttGAN only conducts the reconstruction learning on the generated image conditioned on the original attributes  $a$ , why the preservation ability of attribute-excluding details can be generalized to the generation conditioned on another attributes  $b$ ? We suggest the reason is that, AttGAN transfers the detail preservation ability from the face reconstruction task to the attribute editing task. Since these two tasks share the same input domain and output domain, they are very similar tasks with tiny *transferability gap* [31] between them. Therefore, the detail preservation ability learned from the face reconstruction task can be easily transferred to the attribute editing task. Besides, these two tasks

are learned simultaneously, therefore such transfer is dynamic and the attribute editing learning does not flush the ability of facial detail reconstruction.

### D. Extension for Attribute Style Manipulation

In Sec. III, the attributes are binary represented, i.e., “with” or “without”, which is stiff for real world applications. However, for example, in most cases what one is interested in is adding a certain style of eyeglasses such as sunglasses or thin rim glasses, rather than just with/without eyeglasses. This problem is more difficult because the labeled data with attribute style is unavailable. To enable our AttGAN to manipulate the attribute style, a set of style controllers  $\theta = [\theta_1, \dots, \theta_i, \dots, \theta_n]$  is introduced. Then following [28] and [26], we bind each  $\theta_i$  and the  $i^{th}$  attribute, and maximize the mutual information between the controllers and the output images to make them highly correlated. As a result, such high correlation enables each  $\theta_i$  to control the corresponding attribute of the output images.

As shown in Fig. 3, based on the original AttGAN, we add style controllers  $\theta$  and a style predictor  $Q$ , and the attribute editing is reformulated as

$$x^{\hat{b}} = G_{dec}(G_{enc}(x^a), \theta, b), \quad (16)$$

where  $x^{\hat{b}}$  is expected to not only own the attribute  $b$ , but also be in the style specified by  $\theta$ . According to [28], the mutual information between  $\theta$  and the output images  $x^*$ <sup>4</sup> is obtained by

$$I(\theta; x^*) = \max_Q \mathbb{E}_{\theta \sim p(\theta), x^* \sim p(x^*|\theta)} [\log Q(\theta|x^*)] + const., \quad (17)$$

and is maximized as

$$\max_{G_{enc}, G_{dec}} I(\theta; x^*), \quad (18)$$

where we achieve the mutual information maximization by optimizing the encoder  $G_{enc}$  and decoder  $G_{dec}$ . By correlating the output images with the style controllers via mutual information maximization, AttGAN is able to manipulate the attributes in different styles in a totally unsupervised way.

<sup>4</sup> $x^* \sim G_{dec}(G_{enc}(x^a), \theta, b)$ ,  $x^a \sim p_{data}$ ,  $b \sim p_{attr}$ ,  $\theta_i \sim p_{\theta_i} = \text{Cat}(n_i, \frac{1}{n_i})$ , where  $n_i$  is predefined number of styles for the  $i^{th}$  attribute.

TABLE I  
NETWORK ARCHITECTURES OF ATTGAN FOR  $128 \times 128$  IMAGES.

Encoder ( $G_{enc}$ )	Decoder ( $G_{dec}$ )	Discriminator ( $D$ )	Classifier ( $C$ )
Conv(64,4,2), BN, Leaky ReLU	DeConv(1024,4,2), BN, ReLU	Conv(64,4,2), LN/IN, Leaky ReLU	
Conv(128,4,2), BN, Leaky ReLU	DeConv(512,4,2), BN, ReLU	Conv(128,4,2), LN/IN, Leaky ReLU	
Conv(256,4,2), BN, Leaky ReLU	DeConv(256,4,2), BN, ReLU	Conv(256,4,2), LN/IN, Leaky ReLU	
Conv(512,4,2), BN, Leaky ReLU	DeConv(128,4,2), BN, ReLU	Conv(512,4,2), LN/IN, Leaky ReLU	
Conv(1024,4,2), BN, Leaky ReLU	DeConv(3,4,2), Tanh	Conv(1024,4,2), LN/IN, Leaky ReLU	
		FC(1024), LN/IN, Leaky ReLU	FC(1024), LN/IN, Leaky ReLU
		FC(1)	FC(13), Sigmoid

TABLE II  
NETWORK ARCHITECTURES OF ATTGAN FOR  $64 \times 64$  IMAGES.

Encoder ( $G_{enc}$ )	Decoder ( $G_{dec}$ )	Discriminator ( $D$ )	Classifier ( $C$ )
Conv(64,5,2), BN, Leaky ReLU	DeConv(512,5,2), BN, ReLU	Conv(64,3,1), LN/IN, Leaky ReLU	
Conv(128,5,2), BN, Leaky ReLU	DeConv(256,5,2), BN, ReLU	Conv(64,5,2), LN/IN, Leaky ReLU	
Conv(256,5,2), BN, Leaky ReLU	DeConv(128,5,2), BN, ReLU	Conv(128,5,2), LN/IN, Leaky ReLU	
Conv(512,5,2), BN, Leaky ReLU	DeConv(64,5,2), BN, ReLU	Conv(256,5,2), LN/IN, Leaky ReLU	
	DeConv(3,5,1), Tanh	Conv(512,5,2), LN/IN, Leaky ReLU	
		Conv(512,3,1), LN/IN, Leaky ReLU	
		FC(1024), LN/IN, Leaky ReLU	FC(1024), LN/IN, Leaky ReLU
		FC(1)	FC(13), Sigmoid

\* Conv(d,k,s) and DeConv(d,k,s) denote the convolutional layer and transposed convolutional layer with d as dimension, k as kernel size and s as stride. BN is batch normalization [23], LN is layer normalization [29] and IN is instance normalization [30].

#### IV. IMPLEMENTATION DETAILS

Our AttGAN is implemented by the machine learning system Tensorflow [32] and the code is publicly available at <https://github.com/LynnHo/AttGAN-Tensorflow>. Please refer to the website for more implementation details.

**Network Architecture.** Table I and Table II shows the detailed network architectures of our AttGAN. The discriminator  $D$  is a stack of convolutional layers followed by fully connected layers, and the classifier  $C$  has a similar architecture and shares all convolutional layers with  $D$ . The encoder  $G_{enc}$  is a stack of convolutional layers and the decoder  $G_{dec}$  is a stack of transposed convolutional layers. We also employ the U-Net [33] like symmetric skip connections between the encoder and decoder, which have been shown to produce high quality results on the image translation task [34]. Architectures for  $64 \times 64$  images are used in the comparisons with VAE/GAN [7] and IcGAN [8], and architectures for  $128 \times 128$  images are used in the comparisons with StarGAN [16], Fader Networks [13], Shen et al. [10] and CycleGAN [21].  $384 \times 384$  images are shown in other experiments for better visual effect.

**Training Details.** The model is trained by Adam optimizer [35] ( $\beta_1 = 0.5, \beta_2 = 0.999$ ) with the batch size of 32 and the learning rate of 0.0002. The coefficients for the losses in Eq. (14) and Eq. (15) are set as:  $\lambda_1 = 100, \lambda_2 = 10$ , and  $\lambda_3 = 1$ , which aims to make the loss values be in the same order of magnitude.

#### V. EXPERIMENTS

**Dataset.** We evaluate the proposed AttGAN on CelebA [3] dataset, which contains two hundred thousand images, each of which has annotation of 40 binary attributes (with/without). Thirteen attributes with strong visual impact are chosen in all our experiments, including “Bald”, “Bangs”, “Black

Hair”, “Blond Hair”, “Brown Hair”, “Bushy Eyebrows”, “Eyeglasses”, “Gender”, “Mouth Open”, “Mustache”, “No Beard”, “Pale Skin” and “Age”, which cover most attributes used in the existing works. Officially, CelebA is separated into training set, validation set and testing set. We use the training set and validation set together to train our model while using the testing set for evaluation.

**Methods.** Under the same experimental settings, we compare our AttGAN with two closely related works: VAE/GAN [7] and IcGAN [8]. We also compare AttGAN with the concurrent work StarGAN [16]. All of VAE/GAN, IcGAN, StarGAN and our AttGAN are trained to handle thirteen attributes with a single model. Besides, we compare our AttGAN with the recent Fader Networks [13] (also closely related), Shen et al. [10] and CycleGAN [21]. Shen et al. and CycleGAN can handle only one attribute with one model. Although Fader Networks is capable for multiple attribute editing with one model, in practice, multiple attribute setting makes the results blurry. Therefore, for these three baselines, each attribute has its own specific model. VAE/GAN<sup>5</sup>, IcGAN<sup>6</sup>, StarGAN<sup>7</sup> and Fader Networks<sup>8</sup> are trained by their official code, while Shen et al. and CycleGAN are implemented by ourself.

##### A. Visual Analysis

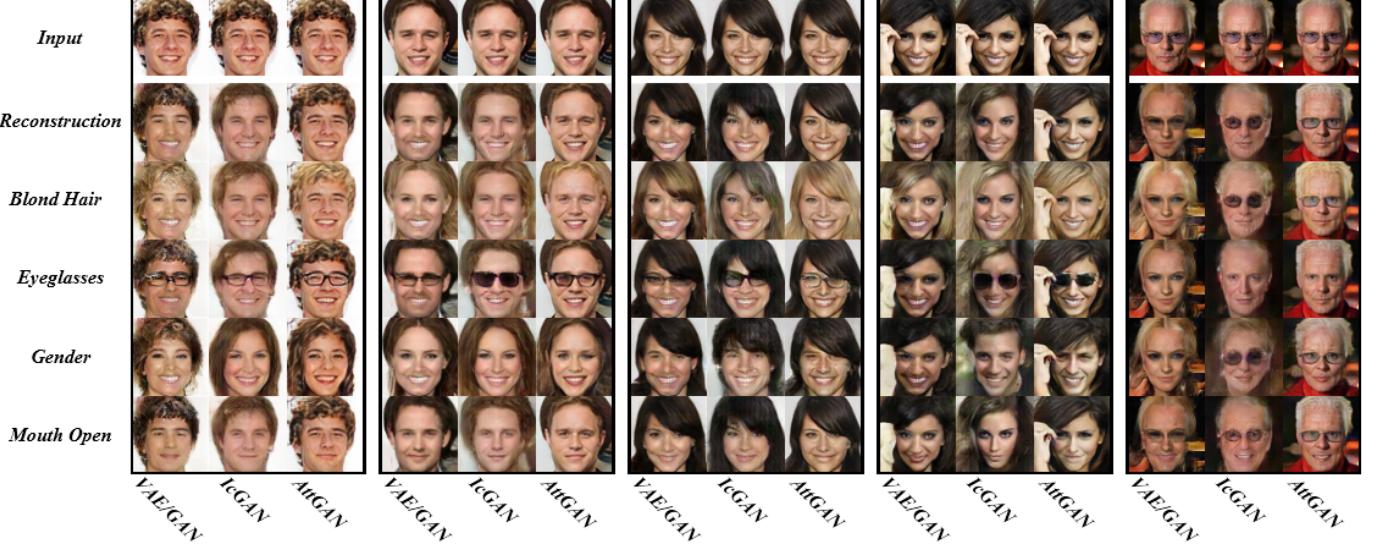
**Single Facial Attribute Editing.** Firstly, we compare the proposed AttGAN with VAE/GAN [7] and IcGAN [8] in terms of single facial attribute editing, shown in Fig. 4a. As can be seen, in some cases VAE/GAN produces unexpected changes of other attributes, for example, all three male inputs become

<sup>5</sup>VAE/GAN: [https://github.com/andersbill/autoencoding\\_beyond\\_pixels](https://github.com/andersbill/autoencoding_beyond_pixels)

<sup>6</sup>IcGAN: <https://github.com/Guim3/IcGAN>

<sup>7</sup>StarGAN: <https://github.com/yunjey/StarGAN>

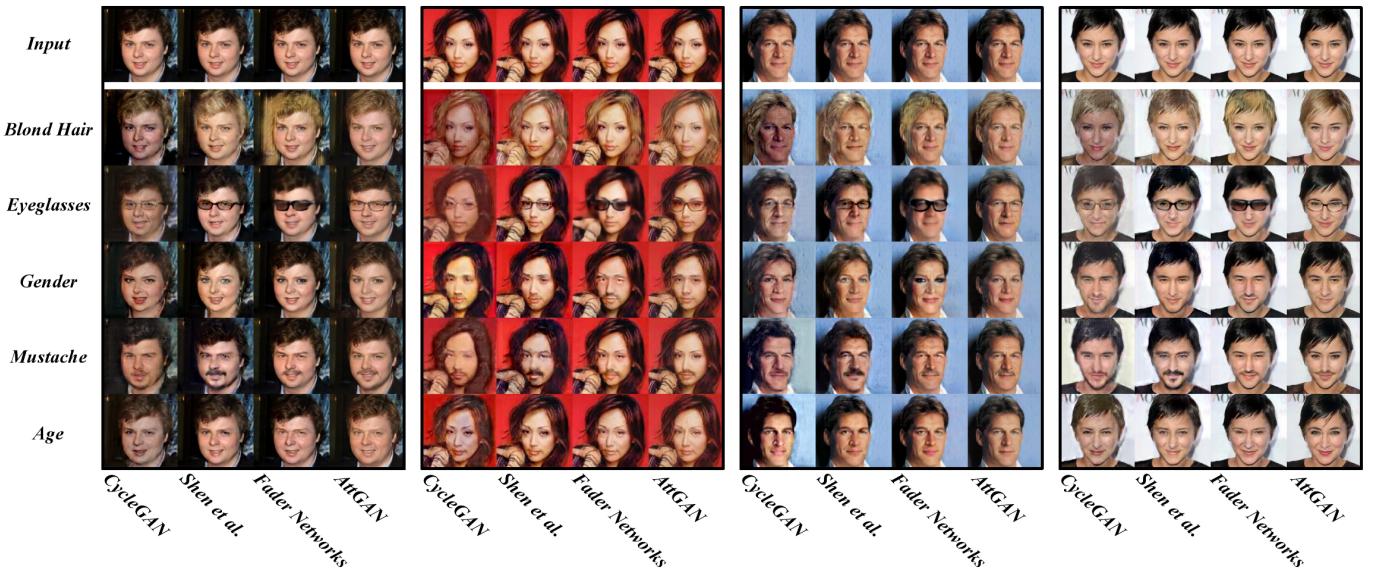
<sup>8</sup>Fader Networks: <https://github.com/facebookresearch/FaderNetworks>



(a) Comparisons with VAE/GAN [7] and IcGAN [8] on editing (inverting) specified attributes.



(b) Comparisons with StarGAN [16] on editing (inverting) specified attributes. Zoom in for better resolution.



(c) Comparisons with Fader Networks [13], Shen et al. [10] and CycleGAN [21] on editing (inverting) specified attributes. Zoom in for better resolution.

Fig. 4. Results of single facial attribute editing. For each specified attribute, the facial attribute editing here is to **invert** it, e.g., to edit female to male, male to female, mouth open to mouth close, and mouth close to mouth open etc.



Fig. 5. Comparisons of multiple facial attribute editing among our AttGAN, VAE/GAN [7] and IcGAN [8]. For each specified attribute combination, the facial attribute editing here is to **invert** each attribute in that combination.



Fig. 6. Illustration of attribute intensity control. Zoom in for better resolution.

female in VAE/GAN when editing the blond hair attribute. This phenomenon happens because the attribute vectors used for editing in VAE/GAN contains highly correlated attributes such as blond hair and female. Therefore, some other unexpected but highly correlated attributes are also involved when using such attribute vectors for editing. IcGAN performs better on accurately editing attributes, however, it seriously changes other attribute-excluding details especially the face identity. This is mainly because IcGAN imposes attribute-independent constraint and normal distribution constraint on the latent representation, which harms its representation ability and results in loss of attribute-excluding information. Compared to VAE/GAN and IcGAN, our AttGAN accurately edits both local attributes (bangs, eyeglasses and mouth open) and global attributes (gender), credited to the attribute classification constraint which guarantees the correct change of the attributes. Moreover, AttGAN well preserves the attribute-excluding details such as face identity, illumination, and background, credited to that 1) the latent representation is constraint free, which guarantees its representation ability for conserving the attribute-excluding information, 2) the reconstruction learning explicitly enable the encoder-decoder to preserve the attribute-excluding details on the generated images.

Comparisons with StarGAN [16] are shown in Fig. 4b. As we can see, both StarGAN and AttGAN accurately edit attributes, but the StarGAN results contain some artifacts while the results of our AttGAN look more natural and realistic.

Comparisons with Fader Networks [13], Shen et al. [10] and CycleGAN [21] are shown in Fig. 4c. The results of Fader Networks especially on adding eyeglasses are blurry, which is very likely caused by the strict attribute-independent constraint on the latent representation. The results of Shen et al. and CycleGAN contain noise and artifacts. Another observation is that, adding “Mustache” makes the female (the second and fourth input in Fig. 4c) become male in Shen et al. and CycleGAN. In the opposite, our AttGAN naturally add the mustache keeping the female’s characteristic well although the model rarely (or never) sees the female with mustache in the training set, which reflects the AttGAN’s superior ability to disentangle attributes (such as male and mustache) and preserve details.

**Multiple Facial Attribute Editing.** All of VAE/GAN [7], IcGAN [8] and our AttGAN can simultaneously edit multiple attributes, and thus we investigate these three methods in terms of multiple facial attribute editing for more comprehensive comparison. Fig. 5 shows the results of simultaneously editing two or three attributes.

Similar to single attribute editing, some generated images from VAE/GAN contain undesired changes of other attributes since VAE/GAN cannot decorrelate highly correlated attributes. As for IcGAN, distortion of face details and over smoothing become even more severe, because its constrained latent representation lead to worse performance in the more complex multiple attribute editing task. By contrast, our

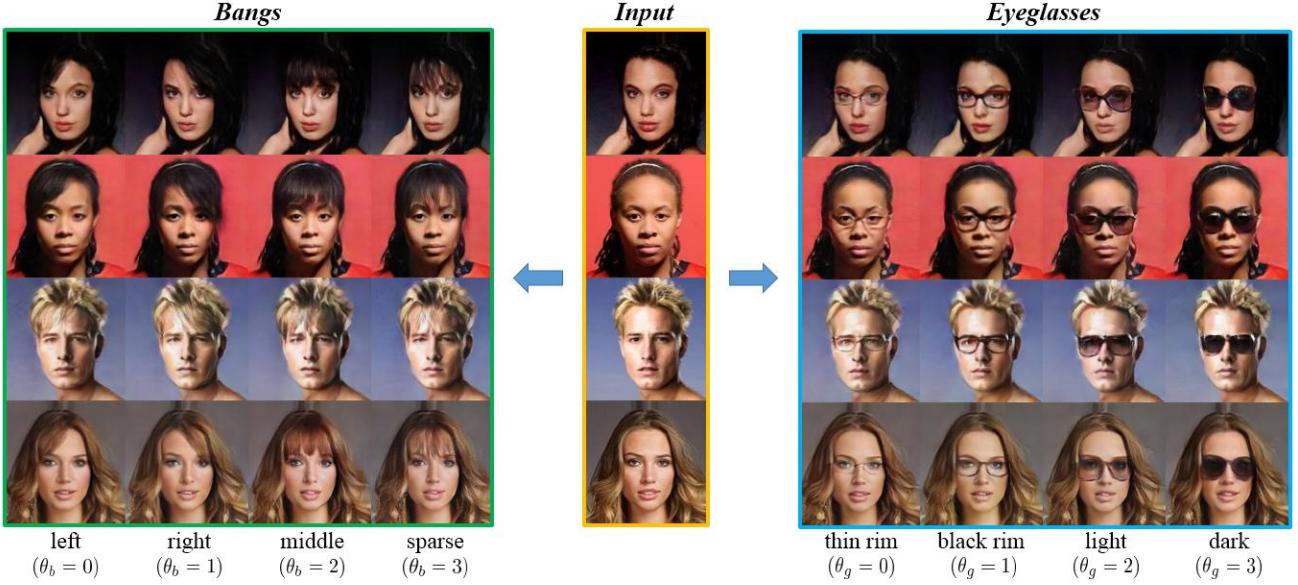
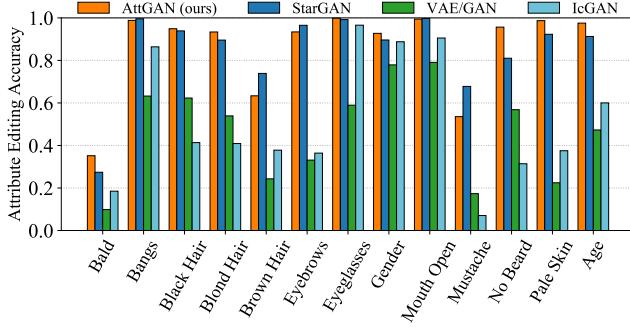
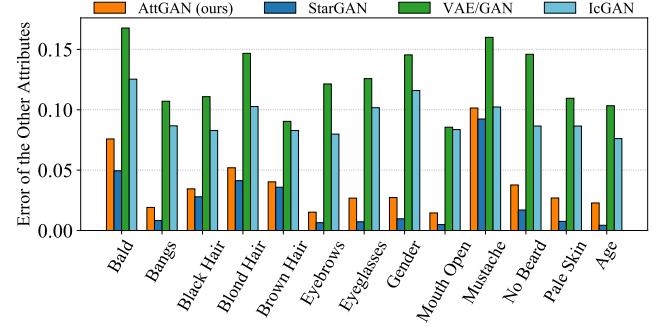


Fig. 7. Exemplar results of attribute style manipulation by using our extended AttGAN.

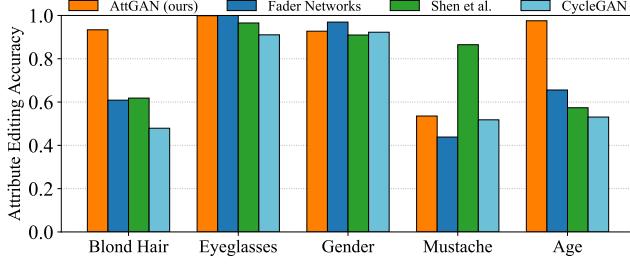


(a) Attribute Editing Accuracy (Higher the Better)

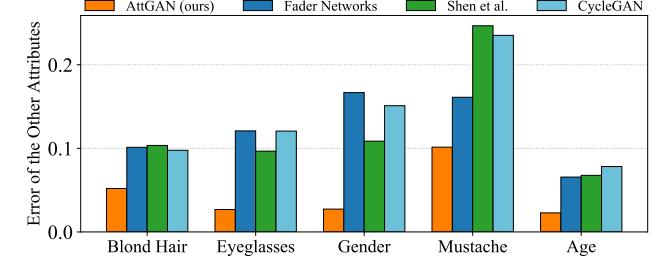


(b) Attribute Preservation Error (Lower the Better)

Fig. 8. Comparisons among StarGAN [16], VAE/GAN [7], IcGAN [8] and our AttGAN in terms of (a) facial attribute editing accuracy and (b) preservation error of the other attributes.



(a) Attribute Editing Accuracy (Higher the Better)



(b) Attribute Preservation Error (Lower the Better)

Fig. 9. Comparisons among Fader Networks [13], Shen et al. [10], CycleGAN [21] and our AttGAN in terms of (a) facial attribute editing accuracy and (b) preservation error of the other attributes.

method still performs well under complex combinations of attributes, benefited from the appropriate modeling of the relation between the attributes and the latent representation.

**Attribute Intensity Control.** Directly applicable for attribute intensity control is a characteristic of our AttGAN. Although AttGAN is trained with binary attribute values (0/1), we find that AttGAN can be generalized for continuous attribute value in testing phase without any modification to its original design. As shown in Fig. 6, with continuous value in  $[0, 1]$  as input, the gradual change of the generated images are smooth and natural.

**Attribute Style manipulation.** Fig. 7 shows the results of the AttGAN extension for attribute style manipulation. As can be seen, different styles of attributes are dug out, such as different sides of bangs: left, right or middle. The extension is quite flexible and allows one to select the style he/she is interested in, rather than a stiff one.

**High Quality Results and Failures.** Fig. 12-14 in supplemental material shows additional results of high quality images with  $384 \times 384$  resolution. Fig. 15 in supplemental material shows some failures. These failures are often caused by the need of large appearance modification, such as editing a face with plenty of hair to “Bald”.

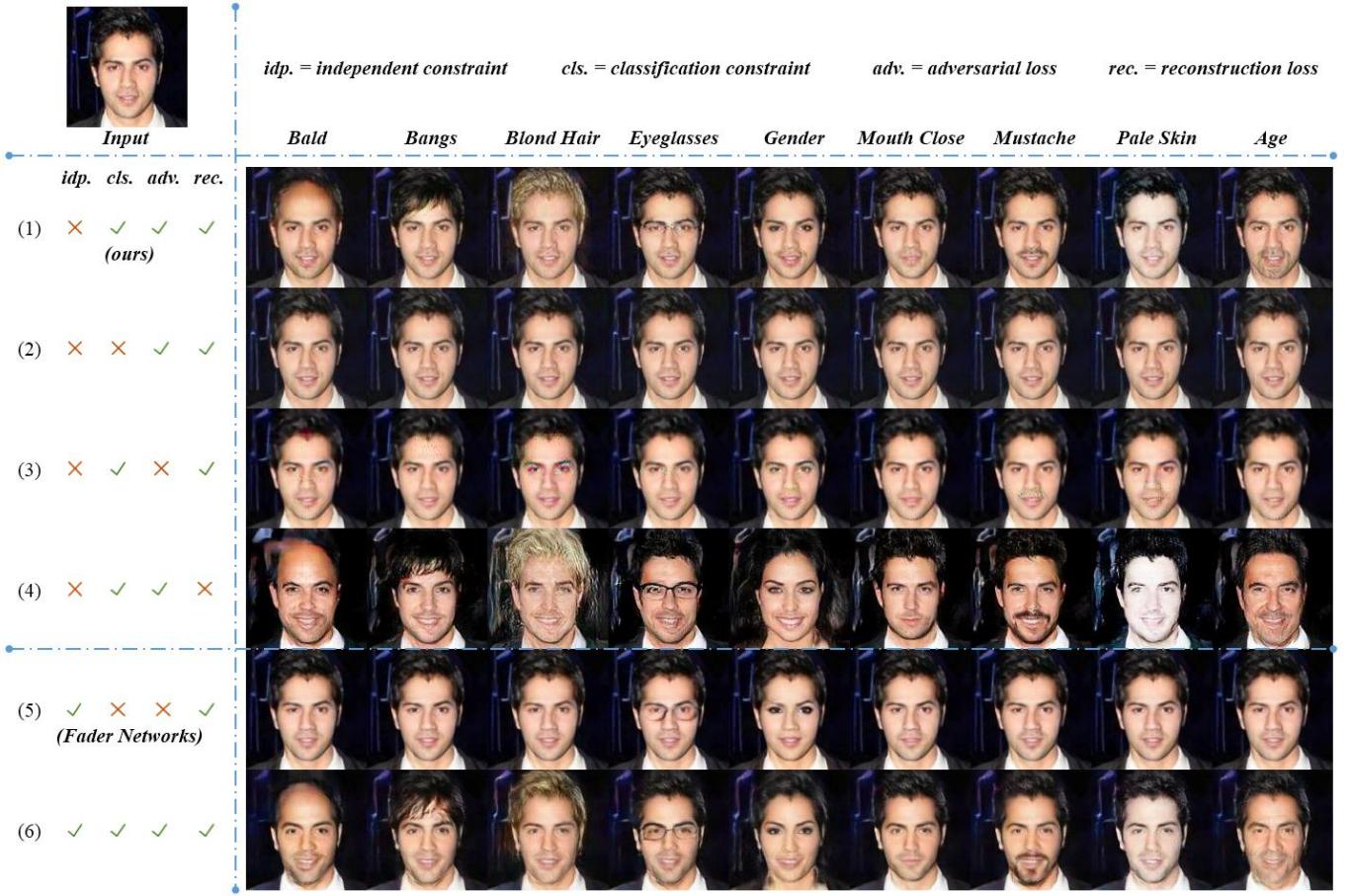


Fig. 10. Effect of different combinations of the four components.

### B. Quantitative Analysis

**Facial Attribute Editing Accuracy/Error.** To evaluate the facial attribute editing accuracy of our AttGAN, an attribute classifier independent of all methods is used to judge the attributes of the generated faces. This attribute classifier is trained on CelebA [3] dataset and achieves average accuracy of 90.89% per attribute on CelebA testing set. If the attribute of a generated image is predicted the same as the desired one by the classifier, it is considered a correct generation, otherwise an incorrect one. Besides, we also evaluate the average preservation error of the other attributes when editing each single attribute.

Fig. 8a shows the attribute editing accuracy of StarGAN [16], VAE/GAN [7], IcGAN [8] and our AttGAN, all of which employ single model for multiple attribute editing. As can be seen, both AttGAN and StarGAN achieve much better accuracy than VAE/GAN and IcGAN, especially on “No Beard”, “Pale Skin” and “Age”. Moreover, the preservation errors of the other attributes of AttGAN and StarGAN are much lower than VAE/GAN and IcGAN as shown in Fig. 8b. As for the comparisons between AttGAN and StarGAN, the attribute editing accuracies of them are comparable, but the attribute preservation error of AttGAN is a bit higher. However, the generated images of our AttGAN are much more natural and realistic than StarGAN (see Fig. 4b).

Furthermore, Fig. 9a and Fig. 9b show the attribute editing accuracy and preservation error of Fader Networks [13], Shen et al. [10] and CycleGAN [21], which employ one specific model for each attribute. As can be seen, all three baselines well edit the attributes which is comparable to AttGAN, but their preservation errors of the other attributes are higher than AttGAN.

### C. Ablation Study: Effect of Each Component

In this part, we evaluate the necessity of the three main components: attribute classification constraint, reconstruction loss and adversarial loss. Besides, we also evaluate the disadvantage of the attribute-independent constraint. In Fig. 10, we show the results of different combinations of these components, where all experiments are based on models which learn to handle multiple attributes with one network. Row (1) contains the results of our AttGAN’s original setting, which are natural and well preserve the attribute-excluding details.

Without the attribute classification constraint (row (2) of Fig. 10), the network just outputs the reconstruction images since there is no signal to force the network to generate the correct attributes. Similar phenomenon (but with some noise) happens when we remove the adversarial loss although the classification constraint is kept (row (3)). One possible reason is that the training with classification constraint but

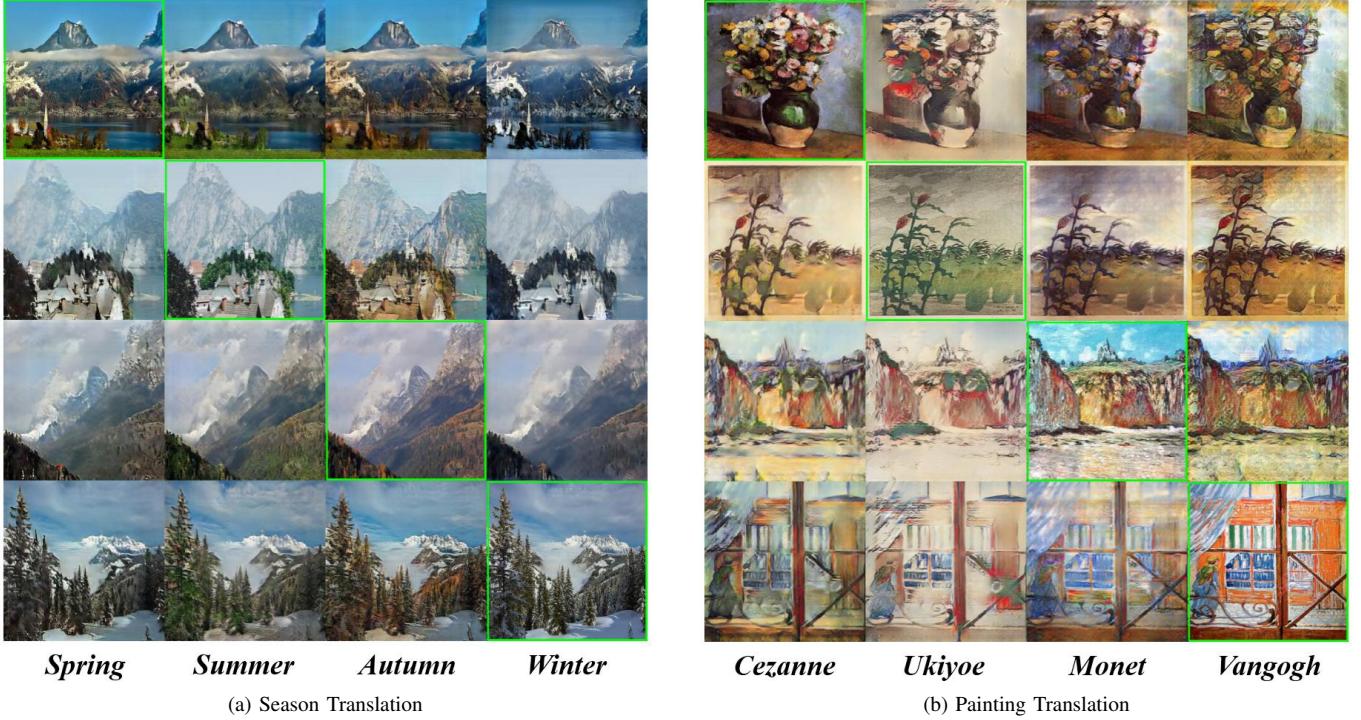


Fig. 11. Exploration of AttGAN on image style translation. The **diagonal** ones are the inputs.

without adversarial loss is similar to making an adversarial attack [36]. Therefore, although the classification constraint exists, the adversarial examples with incorrect attributes still fool the classifier (by the noise). In conclusion, the classification constraint does not work without the adversarial learning, or in other words, the adversarial learning helps to avoid adversarial examples. However, this is another topic needing more theoretical analysis and experiments, which is far beyond this paper.

In row (4) of Fig. 10, we present the results of AttGAN without reconstruction loss. As shown, although the resulting attributes are correct, the face identities change a lot accompanied with many artifacts. Therefore, the reconstruction loss is vital for preserving the attribute-excluding details.

Row (5) of Fig. 10 presents the results of the Fader Networks [13] like setting (attribute-independent constraint + reconstruction learning) and row (6) is AttGAN with attribute-independent constraint. As we can see in the row (5), the Fader Networks like setting works only on eyeglasses, gender and mouth open attributes with unsatisfactory performance. When we combine the AttGAN losses with the Fader Networks losses (row (6)), the attributes is correctly edited but the results contain artifacts and the attribute-excluding details change (e.g., the shape of nose and mouth). These experiments demonstrates that the attribute-independent constraint on the latent representation is not a favorable solution for facial attribute editing, since it constraints the representation ability of the latent code resulting in information loss and degraded output images.

#### D. Exploration of Image Translation

Since facial attribute editing is closely related to image translation, we also try our AttGAN on the image style translation task where we define the style as a kind of attribute. We employ AttGAN on a season dataset [37] and a painting dataset [21] and the results are shown in Fig. 11. As we can see, the results of season are acceptable but the style translation of paintings is not so good accompanied with artifacts and blurriness. Compared to facial attribute editing, image style translation needs more variations on texture and color, a single model might be difficult to simultaneously handle all styles with large variation. However, AttGAN is a potential framework which deserves more explorations and extensions.

## VI. CONCLUSION

From the perspective of facial attribute editing, we reveal and validate the disadvantage of the attribute-independent constraint on the latent representation. Further, we properly consider the relation between the attributes and the latent representation and propose an AttGAN method, which incorporates the attribute classification constraint, the reconstruction learning, and the adversarial learning to form an effective framework for high quality facial attribute editing. Experiments demonstrate that our AttGAN can accurately edit facial attributes, while well preserving the attribute-excluding details, with better visual effect, editing accuracy and lower editing error than the competing methods. Moreover, our AttGAN is directly applicable for attribute intensity control and can be extended for attribute style manipulation, which shows its potential for further exploration.

## ACKNOWLEDGMENT

This work was supported partly by National Key R&D Program of China under contract No.2017YFA0700800, Natural Science Foundation of China under contracts Nos.61390511, 61650202, and 61402443.

## REFERENCES

- [1] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1
- [2] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1
- [3] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 6, 10
- [4] M. Ehrlich, T. J. Shields, T. Almav, and M. R. Amer, "Facial attributes classification using multi-task representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016. 1
- [5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014. 1, 3
- [6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2014. 1, 3
- [7] 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)*, 2016. 1, 3, 6, 7, 8, 9, 10
- [8] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional gans for image editing," in *Advances in Neural Information Processing Systems (NIPS) Workshops*, 2016. 1, 2, 3, 6, 7, 8, 9, 10
- [9] M. Li, W. Zuo, and D. Zhang, "Deep identity-aware transfer of facial attributes," *arXiv preprint arXiv:1610.05586*, 2016. 1, 3
- [10] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3, 6, 7, 8, 9, 10
- [11] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2017. 1, 3
- [12] S. Zhou, T. Xiao, Y. Yang, D. Feng, Q. He, and W. He, "Genegan: Learning object transfiguration and attribute subspace from unpaired data," in *British Machine Vision Conference (BMVC)*, 2017. 1, 3
- [13] 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)*, 2017. 1, 2, 3, 6, 7, 8, 9, 10, 11
- [14] T. Kim, B. Kim, M. Cha, and J. Kim, "Unsupervised visual attribute transfer with reconfigurable generative adversarial networks," *arXiv preprint arXiv:1707.09798*, 2017. 1, 3
- [15] T. Xiao, J. Hong, and J. Ma, "Dna-gan: Learning disentangled representations from multi-attribute images," in *International Conference on Learning Representations (ICLR) Workshops*, 2018. 1, 3
- [16] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 6, 7, 8, 9, 10
- [17] M. Li, W. Zuo, and D. Zhang, "Convolutional network for attribute-driven and identity-preserving human face generation," *arXiv preprint arXiv:1608.06434*, 2016. 3
- [18] P. Upchurch, J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Weinberger, "Deep feature interpolation for image content changes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3
- [19] Y. Bengio, G. Mesnil, Y. Dauphin, and S. Rifai, "Better mixing via deep representations," in *International Conference on Machine Learning (ICML)*, 2013. 3
- [20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. 3, 4
- [21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3, 6, 7, 8, 9, 10, 11
- [22] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations (ICLR)*, 2016. 3
- [23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, 2015. 3, 6
- [24] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," in *International Conference on Machine Learning (ICML)*, 2017. 3, 4
- [25] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems (NIPS)*, 2017. 4, 5
- [26] T. Kaneko, K. Hiramatsu, and K. Kashino, "Generative attribute controller with conditional filtered generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4, 5
- [27] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Advances in Neural Information Processing Systems (NIPS) Workshops*, 2016. 4
- [28] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2016. 4, 5
- [29] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016. 6
- [30] D. U. andAndrea Vedaldi and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016. 6
- [31] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems (NIPS)*, 2014. 5
- [32] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning." 6
- [33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2015. 6
- [34] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6
- [35] D. Kingma and J. Ba, "adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015. 6
- [36] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014. 11
- [37] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, "Combogan: Unrestrained scalability for image domain translation," *arXiv preprint arXiv:1712.06909*, 2017. 11



**Zhenliang He** received the B.E. degree from Beijing University of Posts and Telecommunications and is pursuing the Ph.D. degree from Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. His research interests include pattern recognition, machine learning and computer vision.



**Wangmeng Zuo** (M'09-SM'14) received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007. He is currently a Professor in the School of Computer Science and Technology, Harbin Institute of Technology. His current research interests include image enhancement and restoration, image and face editing, object detection, visual tracking, and image classification. He has published over 70 papers in toptier academic journals and conferences. He has served as a Tutorial Organizer in ECCV 2016, an Associate Editor of the *IET Biometrics and Journal of Electronic Imaging*.

**Meina Kan** is now an Associate Professor with the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), where she received the Ph.D. degree in computer science in 2013. Her research mainly focuses on face detection, face recognition, transfer learning and deep learning.



**Shiguang Shan** is a professor of ICT, CAS, and the deputy director with the Key Laboratory of Intelligent Information Processing, CAS. His research interests cover computer vision, pattern recognition, and machine learning. He has authored more than 200 papers in refereed journals and proceedings in the areas of computer vision and pattern recognition. He was a recipient of the China's State Natural Science Award in 2015, and the China's State S&T Progress Award in 2005 for his research work. He has served as the Area Chair for many international conferences, including ICCV'11, ICPR'12, ACCV'12, FG'13, ICPR'14, and ACCV'16. He is an associate editor of several journals, including the *IEEE Transactions on Image Processing*, the *Computer Vision and Image Understanding*, the *Neurocomputing*, and the *Pattern Recognition Letters*. He is a senior member of the IEEE.



**Xilin Chen** is a professor of ICT, CAS. He has authored one book and more than 200 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He served as an Organizing Committee/Program Committee member for more than 70 conferences. He was a recipient of several awards, including the China's State Natural Science Award in 2015, the China's State S&T Progress Award in 2000, 2003, 2005, and 2012 for his research work. He is currently an associate editor of the *IEEE Transactions on Multimedia*, a leading editor of the *Journal of Computer Science and Technology*, and an associate editor-in-chief of the *Chinese Journal of Computers*. He is a fellow of the China Computer Federation (CCF), IAPR, and the IEEE.





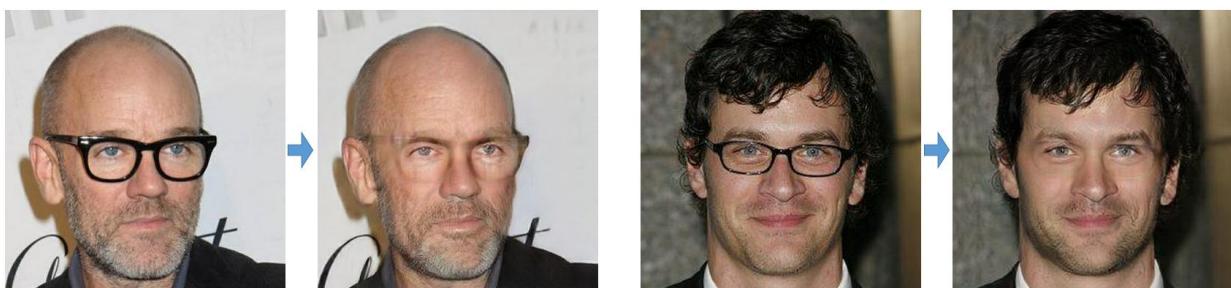
(a) Add Bangs



(b) Remove Bangs



(c) Add Eyeglasses



(d) Remove Eyeglasses



(e) Add Beard



(f) Remove Beard

Fig. 12. Additional AttGAN results of high quality images with  $384 \times 384$  resolution. Zoom in for better resolution.



(a) To Female



(b) To Male



(c) To Black Hair



(d) To Blond Hair



(e) To Bushy Eyebrows + Mouth Open

(f) To Bushy Eyebrows + Mouth Close



(g) To Light Eyebrows + Mouth Open

(h) To Light Eyebrows + Mouth Close

Fig. 13. Additional AttGAN results of high quality images with  $384 \times 384$  resolution. Zoom in for better resolution.



(a) To Male + To Young

(b) To Male + To Old



(c) To Female + To Young

(d) To Female + To Old



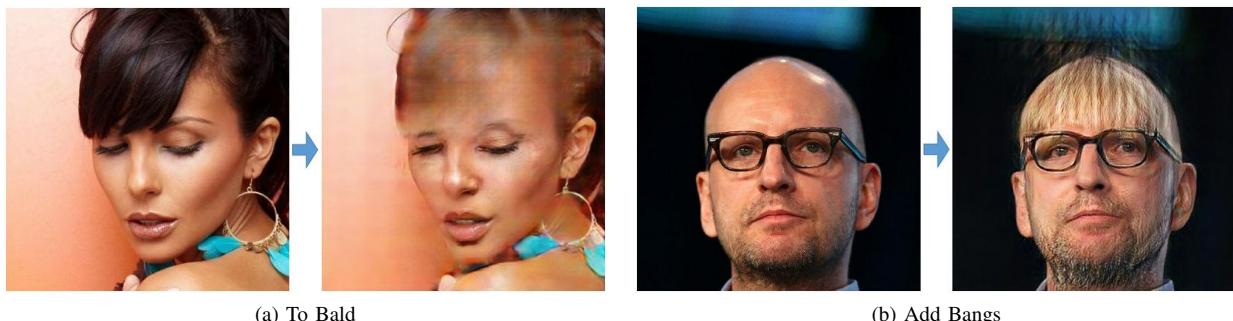
(e) To Blond Hair + Add Beard

(f) To Blond Hair + Remove Beard



(g) To Brown Hair + Add Beard

(h) To Brown Hair + Remove Beard

Fig. 14. Additional AttGAN results of high quality images with  $384 \times 384$  resolution. Zoom in for better resolution.

(a) To Bald

(b) Add Bangs



(c) To Black Hair

(d) Remove Eyeglasses

Fig. 15. Failures, which are often caused by the need of large appearance modification.