ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/321374783

Arbitrary Facial Attribute Editing: Only Change What You Want

Article · November 2017

CITATIONS
22

READS
408

5 authors, including:

Wangmeng Zuo
Harbin Institute of Technology
333 PUBLICATIONS **7,179** CITATIONS
SEE PROFILE

Meina Kan
37 PUBLICATIONS **998** CITATIONS
SEE PROFILE

Shiguang Shan
Chinese Academy of Sciences
301 PUBLICATIONS **10,766** CITATIONS
SEE PROFILE

Xilin Chen
Chinese Academy of Sciences
380 PUBLICATIONS **11,576** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Human Recognition View project

Image compression View project

All content following this page was uploaded by Wangmeng Zuo on 07 December 2017.

The user has requested enhancement of the downloaded file.

arXiv:1711.10678v2 [cs.CV] 6 Dec 2017

Arbitrary Facial Attribute Editing: Only Change What You Want

Zhenliang He[1,2] Wangmeng Zuo[4] Meina Kan[1] Shiguang Shan[1,3] Xilin Chen[1]
[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
[4] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
zhenliang.he@vipl.ict.ac.cn, wmzuo@hit.edu.cn, {kanmeina,sgshan,xlchen}@ict.ac.cn

Abstract

*Facial attribute editing aims to modify either single or multiple attributes on a face image. Since it is practically infeasible to collect images with arbitrarily specified attributes for each person, the generative adversarial net (GAN) and the encoder-decoder architecture are usually incorporated to handle this task. With the encoder-decoder architecture, arbitrary attribute editing can then be conducted by decoding the latent representation of the face image conditioned on the specified attributes. A few existing methods attempt to establish attribute-independent latent representation for arbitrarily changing the attributes. However, since the attributes portray the characteristics of the face image, the attribute-independent constraint on the latent representation is excessive. Such constraint may result in information loss and unexpected distortion on the generated images (*e.g. *over-smoothing), especially for those identifiable attributes such as gender, race* etc. *Instead of imposing the attribute-independent constraint on the latent representation, we introduce an attribute classification constraint on the generated image, just requiring the correct change of the attributes. Meanwhile, reconstruction learning is introduced in order to guarantee the preservation of all other attribute-excluding details on the generated image, and adversarial learning is employed for visually realistic generation. Moreover, our method can be naturally extended to attribute intensity manipulation. Experiments on the CelebA dataset show that our method outperforms the state-of-the-arts on generating realistic attribute editing results with facial details well preserved.*

1. Introduction

This paper investigates the problem of facial attribute editing, which aims to manipulate a face image by controlling the facial attributes of interest (*e.g.* gender, expression, mustache, and age). For conventional face recognition [21, 19] and facial attribute prediction [13, 3] tasks, significant advances have been made along with the development of large scale labeled datasets and deep convolutional neural networks (CNNs). But for facial attribute editing, supervised learning generally is not applicable because it is quite difficult (or even impossible) to collect images of varying attributes for each person. Therefore, one feasible solution to tackle this problem is resorting to the unsupervised methods such as variational autoencoder (VAE) [9] and generative adversarial network (GAN) [4].

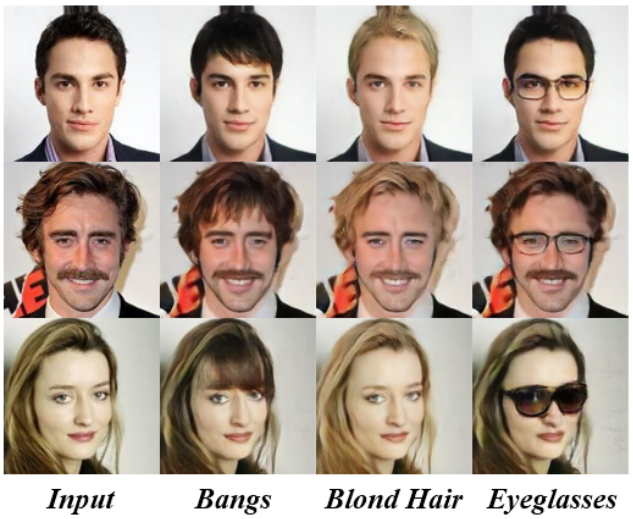

Figure 1. Facial attribute editing results from our AttGAN. The input faces are manipulated to exhibit specified attribute.

Recently, considerable progress has been made on facial attribute editing [11, 17, 12, 20, 23, 10]. However, most existing methods [12, 20, 23] are proposed to train a specific model for each attribute editing subtask. Tak-

ing both the number of attributes and their combinations into account, one has to learn numerous models for handling various attribute editing subtasks, further adding difficulty to their real deployment. Therefore, we focus on arbitrary facial attribute editing which aims to use a single model for any manipulations of facial attributes. In general, the encoder-decoder architecture [11, 17, 10] seems to be a natural choice. Obtaining the latent representation of a face image from the encoder, arbitrary facial attribute editing can be conducted by decoding this latent representation conditioned on specified attributes.

The key issue of arbitrary facial attribute editing based on the encoder-decoder architecture is how to model the relation between the attributes and the latent representation of the face image. For this issue, VAE/GAN [11] represents each attribute as the difference between the mean latent representations of the images with and without this attribute. By adding a single or multiple attribute vectors to the latent representation of a face image, the decoded image from the modified representation is expected to own those attributes. However, such manner cannot separate attribute information from the latent representation and the attribute vector contains highly correlated attributes, making unwanted attributes and artifacts inevitable on the editing results. In IcGAN [17], the latent representation is sampled from normal distribution, which is independent of the attributes. In Fader Networks [10], an adversarial process is introduced to force the latent representation to be invariant to the attributes. However, the attributes portray the characteristics of the face image, meaning that the relation between the attributes and the latent representation is highly complex and difficult to be decoupled. Therefore the attempt to separate all the attribute information from the latent representation is excessive. Simply imposing attribute-independent constraint on the latent representation may result in loss of information and be harmful to the attribute editing, especially for those identifiable attributes such as gender, race *etc*.

In this paper, we present an AttGAN model for arbitrary facial attribute editing reconsidering the relation between the attributes and the latent representation. From the perspective of attribute editing, what we need is not the invariance of the latent representation to the attributes, but just the correct change of the attributes on the generated image. To this end, instead of imposing the independence constraint on the latent representation [17, 10], we adopt an attribute classification constraint on the generated image, just requiring that the attribute manipulation result is correct. In comparison with IcGAN [17] and Fader Networks [10], in our AttGAN, no constraint is imposed on the latent representation, which guarantees its representation ability for further editing.

Besides the attribute classification constraint, AttGAN introduces the reconstruction learning to preserve the attribute-excluding details[1] and only change the specified attributes on the generated image. Moreover, the adversarial learning is employed for the visually realistic generation.

Our 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 can be naturally extended to attribute intensity manipulation and also produces visually plausible results. To sum up, the contribution of this work lies in three folds:

- Properly modeling the relation between the attributes and the latent representation under the principle of just satisfying the attribute editing objective. Our AttGAN removes the strict attribute-independent constraint on the latent representation, and just imposes the attribute classification constraint on the generated image to require the correct change of the attributes.

- Incorporating the reconstruction learning, the adversarial learning and the classification constraint into a unified framework for high quality arbitrary facial attribute editing. The classification constraint guarantees correct attribute editing on the generated image. The reconstruction learning aims at preserving the attribute-excluding details. The adversarial learning is for visually realistic generation.

- Convincing results on arbitrary facial attribute editing. AttGAN outperforms the state-of-the-arts with better perceptual quality for arbitrary facial attribute editing. Moreover, out method can be naturally extended to attribute intensity manipulation.

2. Related Work

Facial Attribute Editing. Li *et al*. [12] present a deep identity-aware attribute transfer (DIAT) model to add/remove an attribute to/from a face image via adversarial learning. Shen and Liu [20] adopt the dual residual learning strategy to simultaneously train two networks for respectively adding and removing a specific attribute. GeneGAN [23] swaps a specific attribute between two given images by recombining the information of their latent representation. These methods [12, 20, 23], however, train different models for the editing of different attributes, which are inapplicable for multiple attribute editing.

Several methods have been proposed for arbitrary facial attribute editing. In VAE/GAN [11], GAN [4] is combined with VAE [9] to learn an latent representation and a generator (or decoder). Attribute editing is conducted by modifying and decoding the latent representation. IcGAN [17]

[1] attribute-excluding details mean all other details of a face image except for the attribute details, such as face identity, illumination and background.

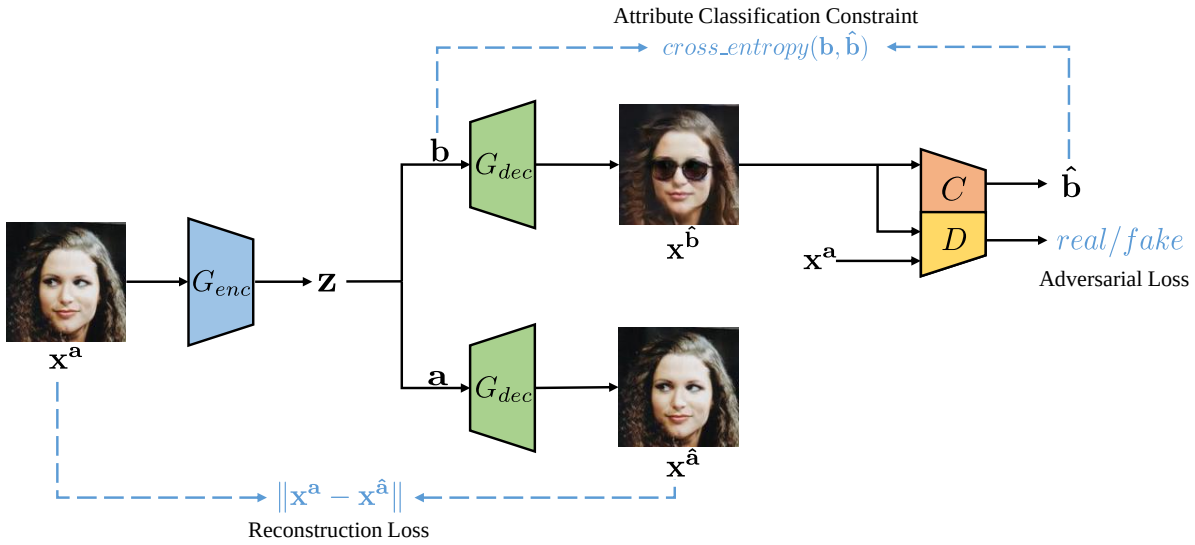

Figure 2. Overview of our AttGAN, which contains three main components, the attribute classification constraint, the reconstruction learning, and the adversarial learning. The attribute classification constraint guarantees correct attribute editing on the generated image. The reconstruction learning aims at preserving the attribute-excluding details. The adversarial learning is for visually realistic generation. G_{enc} in same color is the same one.

separately trains a cGAN [15] and an encoder, requiring that the latent representation is sampled from the uniform distribution and independent of the attributes. Attribute editing in IcGAN is performed by first encoding an image into the latent representation and then decoding the representation conditioned on the given attributes. Fader Networks [10] employs the adversarial learning on the latent representation of an autoencoder to learn attribute invariant representation. Then, the decoder takes the latent representation and arbitrary attribute vector as input to generate the editing result. The attribute-independent constraint on the latent representation of IcGAN and Fader Networks is excessive and harms the representation ability, which may result in information loss and unexpected distortion on the generated images (*e.g.* over-smoothing).

Generative Adversarial Networks. Denoted 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) [4] 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 follow,

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

Theoretically, when the adversarial process reaches the Nash equilibrium, the minimax game attains its global optimum $p(G(\mathbf{z})) \sim p_{data}$ [4]. Subsequently, cGAN [15] and AcGAN [16] have been developed for the conditional generation with given attribute or class label.

GAN is notorious for its unstable training and mode collapse. DCGAN [18] uses CNN and batch normalization [7] for stable training. Subsequently, to avoid mode collapse and further enhance the training stability, WGAN [1] learns to minimize Wasserstein-1 distance between the generated distribution and the real distribution,

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

where D is constrained to be the 1-Lipschitz function implemented by weight clipping. Furthermore, WGAN-GP [5] 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.

3. Attribute GAN (AttGAN)

This section introduces an AttGAN model for arbitrary editing of binary facial attributes, *i.e.* each attribute is represented by 1/0 code for with/without it and all attributes are represented by a vector of 1/0 sequence. In the following, we describe the design principles of AttGAN and introduce the objective for its optimization. Then we present an extension of AttGAN for attribute intensity manipulation.

3.1. Formulation

The architecture of our AttGAN is shown in Fig. 2. In general, it 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. Given a face image $\mathbf{x}^{\mathbf{a}}$ with n binary attributes $\mathbf{a} = (a_1, ..., a_n)$, the encoder G_{enc} is used to encode $\mathbf{x}^{\mathbf{a}}$ into the latent representation, denoted as:

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

Then the process of changing the attributes of $\mathbf{x}^{\mathbf{a}}$ to another attributes $\mathbf{b} = (b_1, ..., b_n)$ is conducted by decoding $\mathbf{z}$ conditioned on $\mathbf{b}$, *i.e.*

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

Thus, the attribute editing problem is formally defined as learning the encoder G_{enc} and decoder G_{dec} which can arbitrarily change the attributes of a real image $\mathbf{x}^{\mathbf{a}}$ to another attributes $\mathbf{b}$. This learning problem is unsupervised, because for each training image $\mathbf{x}^{\mathbf{a}}$ the expected image $\mathbf{x}^{\mathbf{b}}$ with attributes $\mathbf{b}$ is unknown.

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 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 reality.

On the other hand, an eligible attribute editing should only change those desired attributes, while keeping all the other details unchanged. To this end, the reconstruction learning is introduced, making the latent representation $\mathbf{z}$ conserve enough information for recovering the attribute-excluding details. Specifically, for the given $\mathbf{x}^{\mathbf{a}}$, the generated image conditioned on its own attributes $\mathbf{a}$, *i.e.*

$$\mathbf{x}^{\hat{\mathbf{a}}} = G_{dec}(\mathbf{z}, \mathbf{a}), \tag{5}$$

should approximate to $\mathbf{x}^{\mathbf{a}}$ itself.

In summary, the relation between the attributes **a**/**b** and the latent representation $\mathbf{z}$ is implicitly modeled in two sides: (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 $\mathbf{x}^{\hat{\mathbf{a}}}$, which approximates to the input $\mathbf{x}^{\mathbf{a}}$ itself.

Attribute Classification Constraint. From the perspective of facial attribute editing, it is required that the generated image $\mathbf{x}^{\hat{\mathbf{b}}}$ should correctly own the new attributes $\mathbf{b}$. However, there is no ground truth image $\mathbf{x}^{\mathbf{b}}$ to guide the correct generation of $\mathbf{x}^{\hat{\mathbf{b}}}$. Therefore, we deploy 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 follow,

$$\min_{G_{dec}, G_{enc}} \mathcal{L}_{cls_g} = \mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim p_{data}, \mathbf{b} \sim p_{attr}}[\ell_g(\mathbf{x}^{\mathbf{a}}, \mathbf{b})], \tag{6}$$

$$\ell_g(\mathbf{x}^{\mathbf{a}}, \mathbf{b}) = \sum_{i=1}^{n} -b_i \log C_i(\mathbf{x}^{\hat{\mathbf{b}}}) - (1-b_i) \log(1 - C_i(\mathbf{x}^{\hat{\mathbf{b}}})), \tag{7}$$

where p_{data} and p_{attr} indicate distribution of the real images and distribution of the 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 together with the encoder and decoder with the objective below,

$$\min_{C} \mathcal{L}_{cls_c} = \mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim p_{data}}[\ell_r(\mathbf{x}^{\mathbf{a}}, \mathbf{a})], \tag{8}$$

$$\ell_r(\mathbf{x}^{\mathbf{a}}, \mathbf{a}) = \sum_{i=1}^{n} -a_i \log C_i(\mathbf{x}^{\mathbf{a}}) - (1-a_i) \log(1 - C_i(\mathbf{x}^{\mathbf{a}})). \tag{9}$$

Reconstruction Loss. As denoted, the reconstruction learning aims for satisfactory preservation of attribute-excluding details. To this end, the decoder learns 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,

$$\min_{G_{dec}, G_{enc}} \mathcal{L}_{rec} = \mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim p_{data}}[\|\mathbf{x}^{\mathbf{a}} - \mathbf{x}^{\hat{\mathbf{a}}}\|_1], \tag{10}$$

where the ℓ_1 loss is used instead of ℓ_2 loss to suppress the blurriness. The reconstruction learning produces an informative representation, which guarantees the preservation of the attribute-excluding details on the generated image.

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 [1], the adversarial losses for the the discriminator and generator are formulated as below:

$$\min_{\|D\|_L \leq 1} \mathcal{L}_{adv_d} = -\mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim p_{data}} D(\mathbf{x}^{\mathbf{a}}) + \mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim p_{data}, \mathbf{b} \sim p_{attr}} D(\mathbf{x}^{\hat{\mathbf{b}}}), \tag{11}$$

$$\min_{G_{dec}, G_{enc}} \mathcal{L}_{adv_g} = -\mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim p_{data}, \mathbf{b} \sim p_{attr}} D(\mathbf{x}^{\hat{\mathbf{b}}}), \tag{12}$$

where D is the discriminator described in Eq. (2). The adversarial losses above are optimized via WGAN-GP [5].

Overall Objective. By combining the attribute classification constraint, the reconstruction loss and the adversarial loss together, an unified attribute GAN is obtained, which can edit the desired attributes with all the attribute-excluding details well preserved.

Overall, the objective for the encoder and decoder is formulated as follows:

$$\min_{G_{dec}, G_{enc}} \mathcal{L}_{enc\_dec} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{cls_g} + \mathcal{L}_{adv_g}, \tag{13}$$

and the objective for the discriminator and the classifier is formulated as below:

$$\min_{D, C} \mathcal{L}_{dis\_cls} = \lambda_3 \mathcal{L}_{cls_c} + \mathcal{L}_{adv_d}, \tag{14}$$

where the discriminator and the classifier share most layers, λ_1, λ_2 and λ_3 are parameters for balancing different terms.

3.2. Attribute Intensity Manipulation

In the above, the attributes are just considered discretely, *i.e*. "with" or "without" an attribute. But in fact, many attributes usually appear with continuity, such as long hair, medium-length hair, and short hair *etc*. In order to freely manipulate the attribute intensity, we extend our AttGAN as shown in Figure 3(a), where the latent representation $\mathbf{z}$ is divided into two parts, *i.e*. $\mathbf{z}_{int}$ and $\mathbf{z}'$. Thereinto, the functionality of $\mathbf{z}'$ remains the same as $\mathbf{z}$ in Eq. (3), which aims for informative feature representation for arbitrary attribute editing. Additionally, $\mathbf{z}_{int}$ with the same dimension as $\mathbf{a}$/$\mathbf{b}$ is modeled as the attribute intensity controller.

Inspired by the adversarial autoencoder (AAE) [14], we use the adversarial learning to make $\mathbf{z}_{int}$ subject to the multivariate uniform distribution on $[0, 1]$, denoted as $\mathbf{U}(0, 1)$. Different from AAE where the adversarial learning is conducted on the full latent representation, we only imposed the adversarial learning on $\mathbf{z}_{int}$, thus avoiding to harm the representation ability of the latent vector $\mathbf{z}'$. Based on $\mathbf{z}_{int}$, a continuous attribute indicator $\mathbf{b}_{int}$ is obtained as

$$\mathbf{b}_{int} = \mathbf{z}_{int} \cdot (2\mathbf{b} - 1), \tag{15}$$

where $\cdot$ denotes the element-wise product and $\mathbf{b}_{int}$ is therefore subject to $\mathbf{U}(-1, 1)$.

We use $\mathbf{b}_{int}$ as the condition vector for the decoder G_{dec} and the decoding process is re-formulated as below,

$$\mathbf{x}^{\hat{\mathbf{b}}} = G_{dec}(\mathbf{z}', \mathbf{b}_{int}). \tag{16}$$

The whole extended AttGAN can be naturally obtained under the same objective as that in Eq. (13) and Eq. (14). When the optimization converges, each element b_{int_i} in $\mathbf{b}_{int}$ uniformly lies on $[-1, 1]$, and we can slide b_{int_i} on $[-1, 1]$ to control the intensity of the i^{th} attribute, generating images with attributes in different intensity as shown in Figure 3(b) and Figure 8.

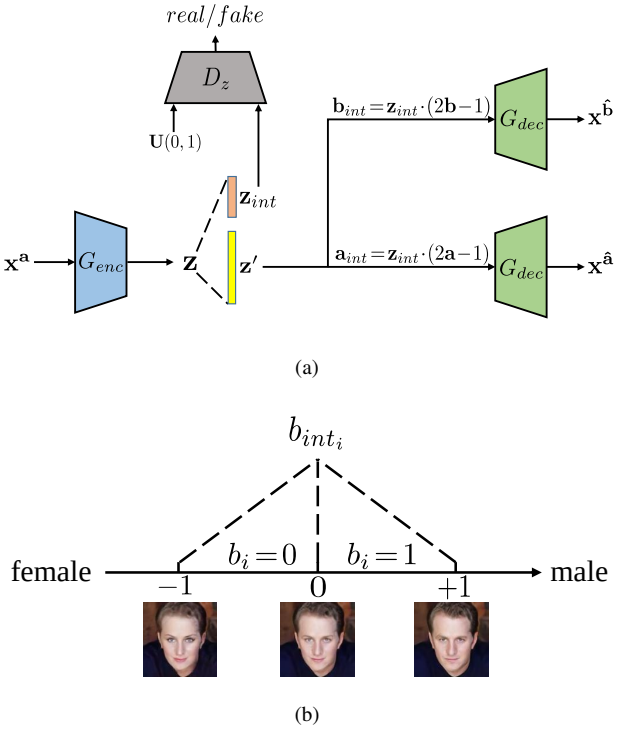

Figure 3. Illustration of AttGAN extension for attribute intensity manipulation. (a) shows the generator (G_{enc} and G_{dec}) based on the re-formulated latent representation and continuous attribute vector, and (b) shows the visual effect of manipulating the gender attribute by sliding b_{int_i}.

4. Experiments

To evaluate the proposed AttGAN, we conduct experiments on the dataset of CelebA [13]. CelebA contains about 200,000 images, each of which has the annotation of 40 binary attributes (with or 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". Officially, CelebA is separated into training set, validation set and testing set. We use the training set and validation set together to train our method while using the testing set for evaluation.

Under the same settings, we compare our AttGAN with two related works: VAE/GAN [11] and IcGAN [17], which are trained with the authors' released codes[2]. Model of 64×64 image is used for comparison with VAE/GAN and IcGAN, while model of 128×128 image is used in the other experiments for better visual effect.

Table 1 and 2 show the detailed network architectures of our AttGAN. The network is trained by using the Adam optimizer [8] ($\beta 1 = 0.5, \beta 2 = 0.999$) with the learning rate of 2×10^{-4}. The coefficients in the loss Eq. (13) and Eq. (14) are set as: $\lambda_1 = 100$, $\lambda_2 = 10$, and $\lambda_3 = 1$.

4.1. Comparison with the Existing Works

Single Facial Attribute Editing. Firstly, we compare the proposed AttGAN with VAE/GAN and IcGAN in terms of single facial attribute editing, as shown in Figure 4. As can be seen, VAE/GAN produces unwanted attributes in some cases, for example, all three male inputs become female in VAE/GAN when editing the blond hair attribute. This phenomenon happens because the attribute vector used for editing in VAE/GAN cannot decouple highly correlated attributes such as blond hair and female. Therefore, some other unwanted but highly correlated attributes are also involved when using such attribute vector 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 distribu-

[2]VAE/GAN: https://github.com/andersbll/autoencoding_beyond_pixels, IcGAN: https://github.com/Guim3/IcGAN

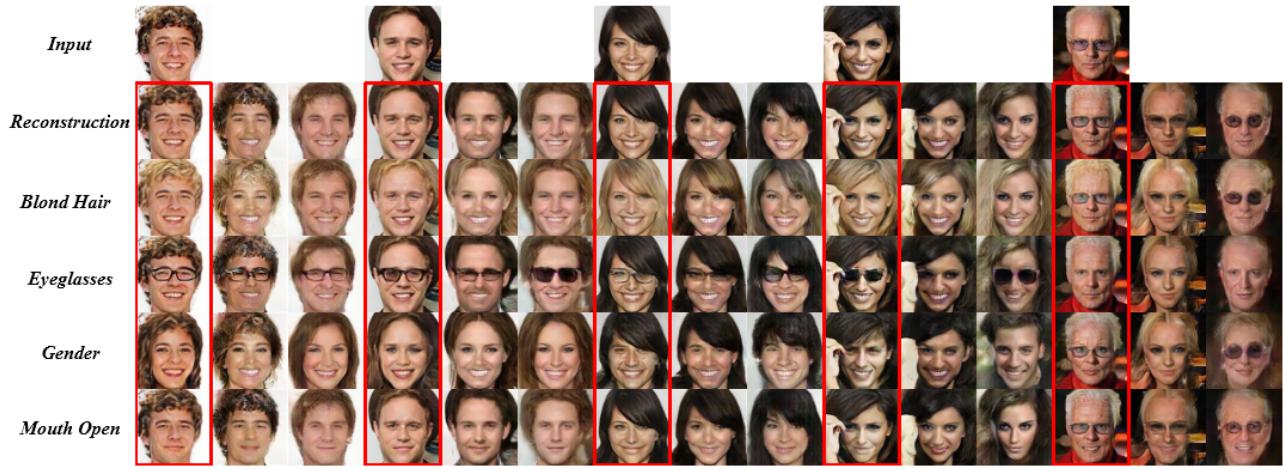

Figure 4. Comparisons of single facial attribute editing among our AttGAN, VAE/GAN and IcGAN. For each input face, the first column with red box are the results of AttGAN , the second and the third columns are the results of VAE/GAN and IcGAN respectively. For each specified attribute, the facial attribute editing here is to invert it, *e.g.* to edit female to male, male to female, mouth open to mouth not open, and mouth not open to mouth open *etc*.

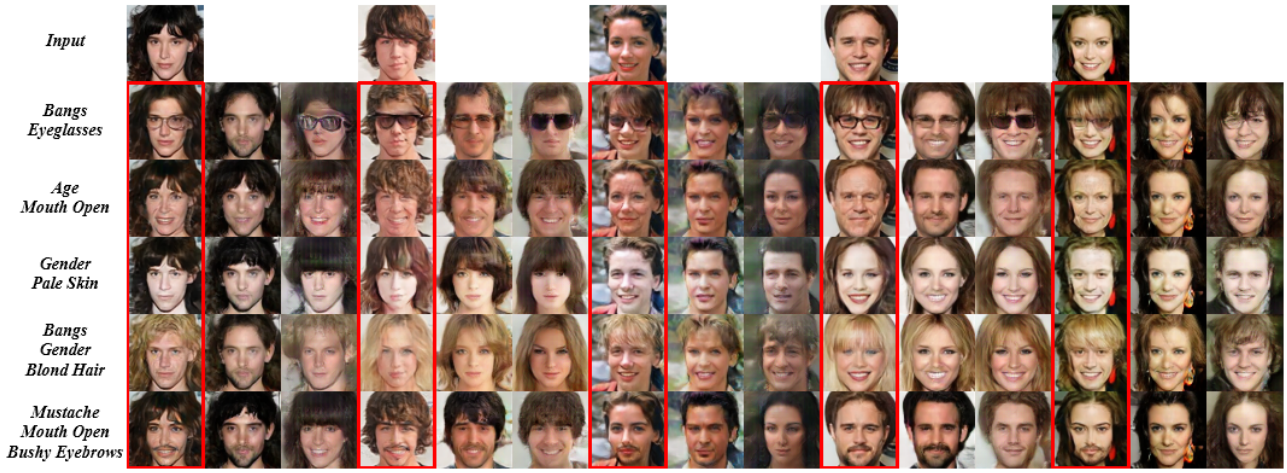

Figure 5. Comparisons of multiple facial attribute editing among our AttGAN, VAE/GAN and IcGAN. For each input face, the first column with red box are the results of AttGAN, the second and the third columns are the results of VAE/GAN and IcGAN respectively. For each specified attribute combination, the facial attribute editing here is to invert each attribute in that combination.

tion 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 as shown in Figure 4. On the one hand AttGAN imposes nearly no constraint on its latent representation, which guarantees its representation ability for conserving the attribute-excluding information. On the other hand, with the help of the reconstruction learning, the encoder and decoder explicitly learn to preserve the attribute-excluding details on the generated images.

Multiple Facial Attribute Editing. All of VAE/GAN, IcGAN 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. Figure 5 shows the results of simultaneously editing two or three attributes.

Similar to the single attribute editing, some generated images from VAE/GAN contain unwanted attributes since VAE/GAN cannot decouple highly correlated attributes. As for IcGAN, distortion of face details and over-smoothing become even more severe, because its constrained latent representation performs worse in the more complex multiple attribute editing task. By contrast, our method still performs well under complex combinations of attributes, benefited from the appropriate modeling of the relation between the attributes and the latent representation.

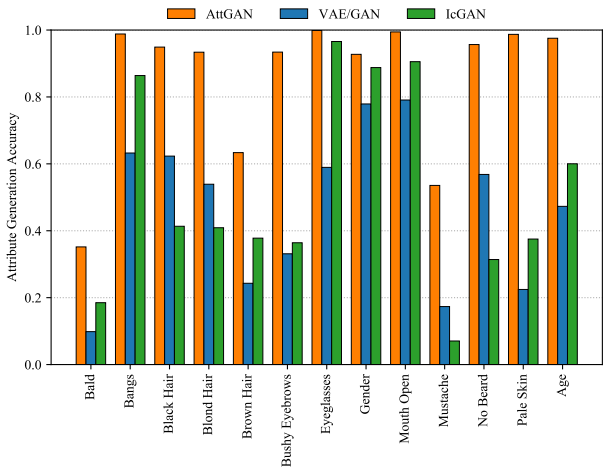

Figure 6. Facial attribute generation accuracy.

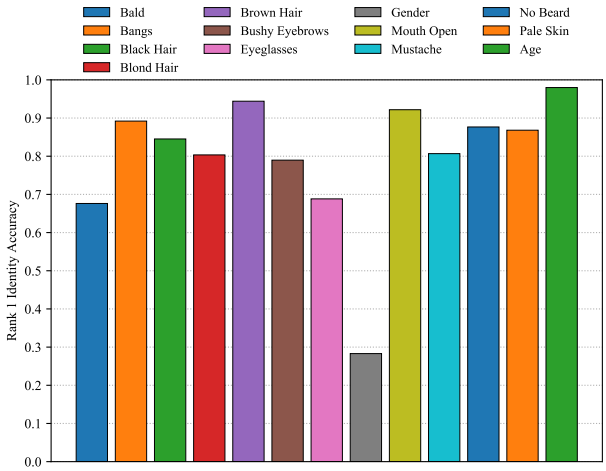

Figure 7. Face identity preservation accuracy of our AttGAN.

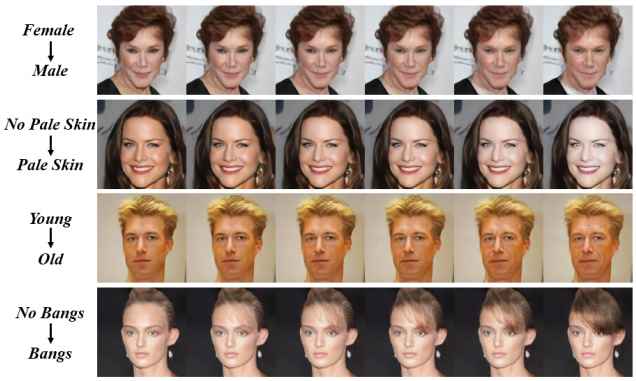

Figure 8. Illustration of the attribute intensity manipulation by using our extended AttGAN.

4.2. Analysis of the Proposed AttGAN

Facial Attribute Generation Accuracy. To evaluate the facial attribute generation accuracy of our AttGAN, a well trained attribute classifier is used to judge the attributes of generated faces. The attribute classifier is trained on CelebA [13] dataset and achieves mean 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. As shown in Figure 6, the generation accuracy of our AttGAN is much better than VAE/GAN and IcGAN, and AttGAN even achieves near-perfect generation for some attributes such as "Bangs", "Eyeglasses", "Mouth Open", and "Pale Skin" *etc*.

Face Identity Preservation. Another aspect for evaluating the facial attribute editing performance is the ability of preserving attribute-excluding details, especially the face identity preservation which is the most important for real application. We train a face recognizer with Inception-ResNet-v2 [22] architecture on the MS-Celeb-1M [6] dataset. The testing set of CelebA [13] is used as the gallery, while the results of attribute editing on the gallery are used as the probe set. The identity preservation ability for editing each attribute by our AttGAN is evaluated by the rank-1 recognition accuracy, shown in Figure 7. As can be seen, the editings of most attributes by our AttGAN well preserve the identity excluding the "Gender" one, since gender is a kind of attribute related to the identity. Thereout, our AttGAN is an effective approach for only changing those attributes what you want.

Attribute Intensity manipulation. To illustrate the intensity manipulation ability of our AttGAN extension, we slide the $\mathbf{b}_{int}$ in Eq. (15) to obtain gradually changing images as shown in Figure 8. As can be seen, the gradual change of the generated images are smooth and natural, benefited from the explicit modeling of intensity manipulation.

Data Augmentation for Attribute Classification. We also employ a well trained AttGAN to augment the CelebA [13] dataset for attribute classification task. Each sample of CelebA training set is augmented to 13 images by editing the 13 chosen attributes respectively. Under a same CNN model, the average attribute accuracies on testing set for the chosen attributes are 93.78% and 94.64% before and after employing the data augmentation, showing that our AttGAN model can be used to benefit the attribute classification task.

Disadvantage of Attribute-Independent Constraint. Both IcGAN [17] and Fader Networks [10] require the latent representation of the face image to be independent of the facial attributes. However, we argue that such constraint is too strict and may harm the attribute editing performance. To validate our argument, like Fader Networks [10], we add

Encoder	Decoder	Discriminator	Classifier
Conv(64,4,2), BN, Leaky ReLU	DeConv(1024,4,2), BN, ReLU	Conv(64,4,2), LN, Leaky ReLU	
Conv(128,4,2), BN, Leaky ReLU	DeConv(512,4,2), BN, ReLU	Conv(128,4,2), LN, Leaky ReLU	
Conv(256,4,2), BN, Leaky ReLU	DeConv(256,4,2), BN, ReLU	Conv(256,4,2), LN, Leaky ReLU	
Conv(512,4,2), BN, Leaky ReLU	DeConv(128,4,2), BN, ReLU	Conv(512,4,2), LN, Leaky ReLU	
Conv(1024,4,2), BN, Leaky ReLU	DeConv(3,4,2), Tanh	Conv(1024,4,2), LN, Leaky ReLU	
		FC(1024), LN, Leaky ReLU	FC(1024), LN, Leaky ReLU
		FC(1)	FC(13), Sigmoid

Table 1. Architecture of our AttGAN with input of size 128×128. 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 [7] and LN is layer normalization [2]

Encoder	Decoder	Discriminator	Classifier
Conv(64,5,2), BN, Leaky ReLU	DeConv(512,5,2), BN, ReLU	Conv(64,3,1), LN, Leaky ReLU	
Conv(128,5,2), BN, Leaky ReLU	DeConv(256,5,2), BN, ReLU	Conv(64,5,2), LN, Leaky ReLU	
Conv(256,5,2), BN, Leaky ReLU	DeConv(128,5,2), BN, ReLU	Conv(128,5,2), LN, Leaky ReLU	
Conv(512,5,2), BN, Leaky ReLU	DeConv(64,5,2), BN, ReLU	Conv(256,5,2), LN, Leaky ReLU	
	DeConv(3,5,1), Tanh	Conv(512,5,2), LN, Leaky ReLU	
		Conv(512,3,1), LN, Leaky ReLU	
		FC(1024), LN, Leaky ReLU	FC(1024), LN, Leaky ReLU
		FC(1)	FC(13), Sigmoid

Table 2. Architecture of our AttGAN with input of size 64×64.

an adversarial process on our AttGAN to enforce the latent representation $\mathbf{z}$ in Eq. (3) to be independent of the attributes. As shown in Figure 9 (c) and (f), the degradation and over-smoothing happen after imposing such attribute-independent constraint, as such excessive constraint weakens the representation ability of the latent representation. Furthermore, when we remove the attribute classification constraint and only impose the attribute-independent constraint on our model, the editing results become even worse as shown in Figure 9 (d) and (g). These experiment results illustrate that the attribute-independent constraint on the latent representation is too strict and harms the editing, while our attribute classification constraint effectively benefits the attribute editing.

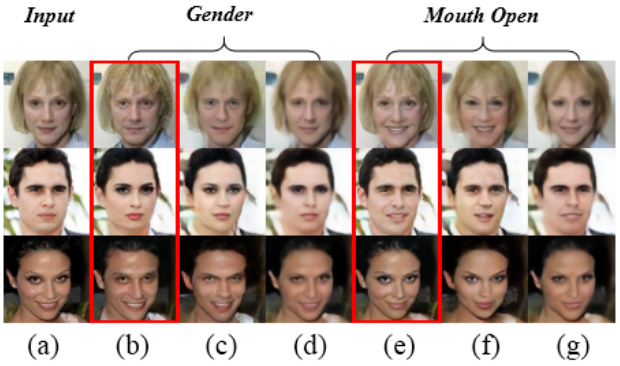

Figure 9. Results of AttGAN with and without attribute-independent constraint. Columns of (b) and (e) are the results of our original AttGAN, (c) and (f) are the results of AttGAN with attribute-independent constraint, and (d) and (g) are the results of AttGAN with attribute-independent constraint and without the attribute classification constraint.

5. Conclusion and Future Work

From the perspective of facial attribute editing, we demonstrate the disadvantage of the attribute-independent constraint on the latent representation, and we propose an AttGAN model properly modeling the relation between the attributes and the latent representation for more favorable facial editing which is able to only change those attributes what you want. Our AttGAN incorporates the reconstruction learning, the adversarial learning and the attribute classification constraint together, forming an end-to-end network for arbitrary facial editing. Furthermore, we extend our AttGAN to attribute intensity manipulation. Experiments demonstrate that our AttGAN can accurately edit facial attributes while well preserving the attribute-excluding details, with much better visual effect than the related works.

Facial attribute style manipulation such as modifying the style of the eyeglasses is highly correlated with facial attribute editing task, which concentrates on the finer facial detail editing and is also an interesting problem. In future work, we will concentrate on the research of the facial attribute style manipulation.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3, 4

[2] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 8

[3] M. Ehrlich, T. J. Shields, T. Almaev, 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

[4] 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, 2, 3

[5] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017. 3, 4

[6] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *European Conference on Computer Vision (ECCV)*, 2016. 7

[7] 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, 8

[8] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[9] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[10] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato. Fader networks: Manipulating images by sliding attributes. *arXiv preprint arXiv:1706.00409*, 2017. 1, 2, 3, 7

[11] 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, 2, 5

[12] M. Li, W. Zuo, and D. Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016. 1, 2

[13] 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, 5, 7

[14] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 5

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

[16] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016. 3

[17] 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, 5, 7

[18] 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

[19] 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

[20] W. Shen and R. Liu. Learning residual images for face attribute manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[21] 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

[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 7

[23] S. Zhou, T. Xiao, Y. Yang, D. Feng, Q. He, and W. He. Genegan: Learning object transfiguration and attribute subspace from unpaired data. *arXiv preprint arXiv:1705.04932*, 2017. 1, 2

View publication stats