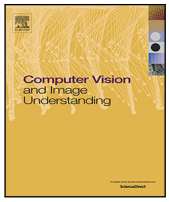
Computer Vision and Image Understanding 217 (2022) 103347

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Progressive editing with stacked Generative Adversarial Network for multiple facial attribute editing

Patrick P.K. Chan *, Xiaotian Wang, Zhe Lin, Daniel S. Yeung

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

ARTICLE INFO

Communicated by Nikos Paragios

Keywords:
Facial multiple attribute editing
Generative Adversarial Networks
Imbalance problem

ABSTRACT

Generative Adversarial Network (GAN) based facial attribute editing has been successfully applied to many real world applications. However, most of existing methods suffer from semantic entanglement and imprecise editing when handling multiple facial attributes. The situation is worse when the samples with minority attribute values are insufficient, and majority attribute values dominate the learning easily. A stacked conditional GAN (cGAN) is proposed in this study aiming at solving these problems. Multiple attribute editing is broken down into several single attribute editing tasks which have been learned by base cGANs individually. Moreover, samples with a minority attribute value are paid more attention in learning. This proposed method not only reduces the difficulty of multiple attribute editing but also mitigates the imbalance problem. The residual image learning is applied to our model to reduce the difficulty of the image generation. The superiority of our model is demonstrated and compared with popular GAN-based facial attribute editing methods experimentally in terms of image quality, editing accuracy and training cost. The results confirm that our proposed model outperforms the other methods, especially in imbalance situations.

1. Introduction

Facial attribute editing manipulates particular attributes of a face image, for example, mustache, gender, skin color and eyeglasses. This visual task achieves semantical controllability and fine-grained image translation for face images. Facial attribute editing has been applied to many applications including data augmentation (Zhang et al., 2018b), web glow (Wang et al., 2016), the synthesis of virtual character in a movie (Zhang et al., 2018a).

Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) achieves satisfying performance in facial attribute editing. Given a dataset containing face images with different attribute labels, the trained generator of GAN manipulates attributes of an image, while the discriminator aims to distinguish the original and edited images. The quality of an edited image is improved by competition between the generator and discriminator of GAN. However, GAN (Li et al., 2016) focuses on a single operation, *i.e.* either adding or removing, on one attribute. It may not handle a multi-attribute editing problem efficiently.

A bidirection mapping on multiple attributes is considered in some studies. Conditional GAN (cGAN)-based methods (Mirza et al., 2016; Choi et al., 2018; He et al., 2019) manipulate attributes of a face image according to an attribute label vector indicating which attributes should be added or removed, while multiple attributes are edited by transforming between an image and a latent code representing presence of attributes in representation-based GAN (rGAN)-based methods (Perarnau et al., 2016).

The multi-attribute facial editing is a tough learning problem due to a huge number of attribute combinations, *i.e.* there is a number of attributes and each attribute may contains a number of values. Similar to other machine learning problems, the collected samples are costly and usually limited. They cannot represent all attribute combinations completely. The current multi-attribute facial editing methods suffer from three issues illustrated by StarGAN, AttGAN and IcGAN in Fig. 1. *Attribute entanglement* indicates manipulating one attribute usually causes the change on other attributes. Although only the mustache is the target attribute, the mouth is also changed in the generated images. *Imprecise editing* including undesired flaws appears in the nose, eyes and hair in the generated images. In addition, the current methods have *bad performance on minority attribute values*. For example, the glasses are not removed completely. Minority samples are ignored easily since traditional learning methods treat samples equally. Moreover, the single model for multiple attributes may cause the failure on the balance of all attributes due to the interference.

This study aims to tackle the problems mentioned above by proposing Progressive Editing with Stacked Generative Adversarial Network, referred as PESGAN. Our model simplifies the multi-attribute editing

* Corresponding author.
E-mail address: patrickchan@ieee.org (P.P.K. Chan).

https://doi.org/10.1016/j.cviu.2021.103347
Received 8 June 2021; Received in revised form 12 December 2021; Accepted 27 December 2021
Available online 31 December 2021

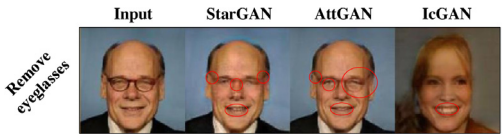


Fig. 1. The problems of existing facial multi-attribute editing illustrated by StarGAN, AttGAN, and IcGAN.

problem by breaking down it into a number of single-attribute problems. One cGAN is trained independently for each attribute to generate a residual image. The final image is generated by stacking the source image and generated residual images in the pixel level.

As the ratio between the sizes of the combinations containing minority attributes and combinations containing majority attributes may increase significantly in multiple-attribute manipulation, considering that single-attribute problems can avoid the aggravation of imbalance problems. Moreover, the imbalance problem is further tackled by assigning a larger weight to the samples with a minority attribute value to punish poor performance on minority samples. As a result, the minority attribute editing can be learnt better in our model. In addition, one advantage of our model is flexibility. Re-training the whole model is not necessary for considering an additional feature or ignoring an existing feature, but only adding or removing a cGAN is needed. As a simpler learning problem requires a smaller network structure, the training time of our divide and conquer method is expected to be lower than traditional multi-attribute editing methods. One possible drawback of our cascade framework is the error accumulation. As manipulating a single attribute is much simpler than manipulating multiple ones, a small error is expected to be made in each manipulation. As a result, the accumulated error of a single attribute manipulation in our method may not necessarily be larger than the error made by multiple-attribute manipulation methods. Our proposed model is experimentally evaluated and compared with the state-of-arts by qualitative and quantitative evaluation for editing single and multiple facial attributes using CelebA and RaFD dataset. The experimental results confirm that the quality of images generated by our model is the best, especially in imbalanced situations. The ablation study is also included to validate the effectiveness of main components of our method.

The rest of the paper is organized as follows. Section 2 introduces facial attribute editing and its methods. Our proposed model is devised and discussed in Section 3. In Section 4, the experimental results and discussion are given. Finally we conclude our work and discuss future work in Section 5.

2. Background

The concepts related to this study, including GANs and facial attribute editing, are introduced in this section.

2.1. Generative Adversarial Networks (GANs)

GANs have drawn significant attention recently, and achieved satisfying results in various tasks, for example, image-to-image translation (Isola et al., 2017; Zhu et al., 2017), data augmentation (Frid-Adar et al., 2018; Pfeiffer et al., 2019) and representation learning (Tran et al., 2017; Li et al., 2018a,b). A generative problem is formulated as an adversarial game between a discriminator and generator. The discriminator distinguishes the real data $\mathbf{x}$ from the fake one while the generator seeks to synthesize realistic data $G(\mathbf{z})$ from the input $\mathbf{z}$ to fool the discriminator. The minimax objective endows the generator effective ability.

Conditional GANs (cGANs) is introduced to generate an image according to the preference of users represented by attribute vector. It enhances flexibility of GAN significantly. The stacked structure (Zhang et al., 2017; Huang et al., 2017; Xu et al., 2018) is one of the efficient skills for synthesizing high-resolution image. An image is input to sub-models sequentially. The final image is generated by the last sub-model. In previous works, each sub-model aims to improve the image quality. Moreover, all sub-models are trained together. Our study applies the framework of stacked GAN to multiple facial attribute editing with independent training strategy for sub-models.

2.2. Facial attribute editing

Facial attribute editing aims to modify characteristics of a face image. GANs are one of the common and useful techniques which yields impressive synthesized image with high quality. The methods are mainly divided into two categories: single (Li et al., 2016) and multi-attribute (Choi et al., 2018) editing methods. For single attribute editing, DIAT designs two identity-related loss components to control an one-way modification on a facial attribute, while CycleGAN, Residual GAN (Shen et al., 2017) and SPA-GAN (Emami et al., 2020) achieve bi-direction mapping on an attribute. Moreover, residual image learning is incorporated in Residual GAN for facilitating model in order to focus on the attribute region with modest modification on all pixels. It reduces the difficulty of the image generation.

Multi-attribute editing methods are usually implemented by cGAN or rGAN. An attribute label vector indicates the attribute modification of a face image. Attribute classification and cycle-consistent loss are used in StarGAN, which allows simultaneous learning on different attributes in cGAN. AttGAN argues that the classification loss may be too aggressive to maintain attribute-irrelevant information. An unique attribute classification constraint is proposed to enhance the attribute-irrelevant information. On the other hand, rGAN-based methods determine an explicit attribute representation vector. IcGAN trains a pair of an encoder and decoder by regression and adversarial learning. The encoder generates the attribute representation vector and attribute-irrelevant representation vector, while the decoder recovers an image from the vectors.

3. Progressive Editing with Stacked Generative Adversarial Network (PESGAN)

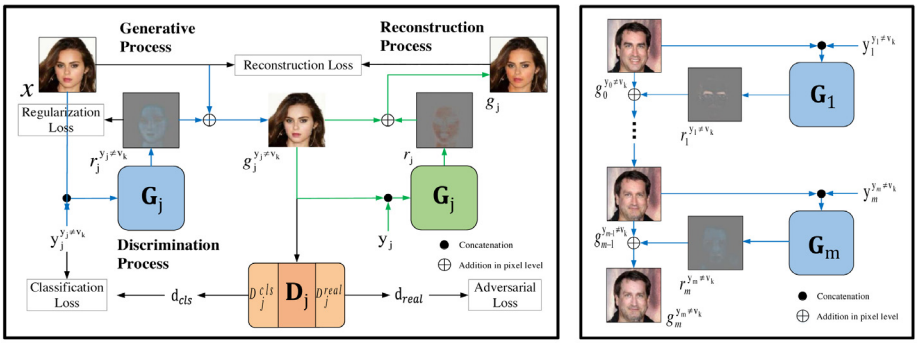
This section presents the framework of our proposed Progressive Editing with Stacked Generative Adversarial Network (PESGAN) for facial multi-attribute editing. A stack structure used in our model avoids *the attribute entanglement* by manipulating attributes separately. Moreover, an image is generated by using residual images, which can edit attributes precisely (Shen et al., 2017), to reduce the accumulated error of the stack structure. Considering that single-attribute problems can avoid the aggravation of imbalance problems. Furthermore, samples with a minority attribute value are emphasized in training by assigning a large weight. The better performance on the editing of minority attribute values is expected.

3.1. Framework

Given a dataset $\mathcal{D}$ containing n face images $(\mathbf{x}, \mathbf{y})$, where $i = 1, 2, \ldots, n$. $\mathbf{x} \in \mathcal{X}$ is a feature vector representing the pixels of the ith image, while $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ is the attribute vector, where $y_j \in \mathcal{Y}_j$ indicates the value of the jth attribute, and $|\mathcal{Y}_j|$ denotes the number of different values of the jth attribute.

cGAN_j is trained independently to manipulate the jth attribute y_j, where $j = 1, 2, \ldots, m$. The generator and discriminator of cGAN_j, denoted by G_j and D_j, are trained simultaneously under the framework of eGAN. G_j aims to minimize four weighted loss functions, including adversarial loss, classification loss, reconstruction loss and regularization loss, while adversarial loss and classification loss are minimized by D_j. The training framework is illustrated in Fig. 2(a).

During the image generation phase, a face image $\mathbf{x}$ and target editing attribute vector $\mathbf{y}' = (y_1^{y_1 \neq v_k}, y_2^{y_2 \neq v_k}, \ldots, y_m^{y_m \neq v_k})$ are given, where



(a) Generator Training (b) Image Generation

Fig. 2. Block diagrams of our proposed model.

v_k indicates the value of the jth attribute k. Let $\mathbf{g}_0^{y_0 \neq v_k}$ be $\mathbf{x}$. For each G_j, $\mathbf{g}_{j-1}^{y_{j-1} \neq v_k}$ together with the jth target attribute value $\mathbf{y}_j^{y_j \neq v_k}$ are used to calculate $\mathbf{r}_j^{y_j \neq v_k}$ by G_j, where $j = 1, 2, \ldots, m$. $\mathbf{r}_j^{y_j \neq v_k}$ is then combined with $\mathbf{g}_{j-1}^{y_{j-1} \neq v_k}$ using the element-wise addition as a complete image $\mathbf{g}_j^{y_j \neq v_k}$. Finally, $\mathbf{g}_m^{y_m \neq v_k}$ is generated after editing m attributes of $\mathbf{x}$ according to $\mathbf{y}'$. The image generation is illustrated in Fig. 2(b).

3.2. Training process

Each cGAN of our model is trained independently for editing one attribute. The training of the generator G_j and discriminator D_j of cGAN$_j$ under the framework of cGAN (Mirza et al., 2016) is described in this section. A training set $\mathcal{D}_j$ is generated by containing all images and their jth attribute label from $\mathcal{D}$, *i.e.* $\mathcal{D}_{G_j} = \{(\mathbf{x}, y_j) | (\mathbf{x}, \mathbf{y}) \in \mathcal{D}$ and $\mathbf{y} = (y_1, y_2, \ldots, y_m)\}$. For each x, a target attribute value $y_j^{y_j \neq v_k} \in \mathcal{Y}_j'$ which is different from y_j and chosen randomly for training.

G_j generates the target residual image with the modified jth attribute $\mathbf{r}_j' \in \mathcal{R}_j'$ according to $\mathbf{x}$ and y_j', *i.e.* $\mathbf{r}_j^{y_j \neq v_k} = G_j(\mathbf{x}, y_j^{y_j \neq v_k})$. The whole image $\mathbf{g}_j^{y_j \neq v_k} \in \mathcal{G}_j'$ is formed by adding $\mathbf{x}$ and $\mathbf{r}_j^{y_j \neq v_k}$. Two discriminators D_j^{cls} and D_j^{real} are trained to verify whether a generated image is realistic and the attributes are edited appropriately. D_j^{cls} and D_j^{real} share the same convolutional network but use different fully connected layers.

3.2.1. Attention on minority samples

As minority samples may be ignored easily in an imbalance problem, each sample in $\mathcal{D}_{G_j}$ is assigned a weight w representing its importance in training. Samples with a minority attribute value are assigned a larger value, vice versa. For the sample $\mathbf{x}$, w is calculated according to the value of the jth attribute (v_k), defined as follows:

$$w = \max_{i \in \{1, \ldots, |Y_j|\}} (n_{y_j = i}) / n_{y_j = v_k}, \tag{1}$$

where $|Y_j|$ denotes the number of different values of the jth attribute and $n_{y_j = v_k}$ represents the number of samples whose the jth attribute is k, *i.e.* $n_{y_j = v_k} = |\{(\mathbf{x}, y) | (\mathbf{x}, y) \in D_{G_j}$ and $y_j = v_k\}|$, and $||$ is the cardinality of a set. w measures the ratio between the largest number of samples with the same value in the jth attribute and the number of samples with v_k in the jth. This ratio is greater than or equal to 1. Smaller $n_{y_j = v_k}$ indicates larger w, vice versa. The weight encourages the learning focus more on the samples with a minority attribute value.

3.2.2. Generator training

The objective function of the generator G_j in our model containing four components, *i.e.* adversarial, classification, reconstruction and regularization losses, is defined as:

$$\min_{G_j} \mathcal{L}^G = \mathcal{L}_{adv}^G + \lambda_{cls}^G * \mathcal{L}_{cls}^G + \lambda_{rec}^G * \mathcal{L}_{rec}^G + \lambda_{l1}^G * \mathcal{L}_{l1}^G \tag{2}$$

where λ_{cls}^G, λ_{rec}^G, λ_{l1}^G are tradeoff parameters, indicating importance of corresponding loss. One should be noted that all these four loss components considers the minority of samples in calculation. These loss components are introduced in detail as follows.

Adversarial Loss quantifies the realistic level of a generated image. It is defined as:

$$\mathcal{L}_{adv}^G = \mathbb{E}_{(y_j, \mathbf{g}_j^{y_j \neq v_k}) \sim (\mathcal{Y}_j, \mathcal{G}_j')} [w * log(1 - D_j^{real}(\mathbf{g}_j^{y_j \neq v_k}))], \tag{3}$$

where $D_j^{real}(\cdot)$ is a discriminator aiming to judge the realness of input image. The similarity between the real image and image synthesized by G_j increases by minimizing the adversarial loss.

Classification Loss measures whether the jth attribute of $\mathbf{x}$ is edited corresponding to its target value $y_j^{y_j \neq v_k}$. This loss adopts weighted binary cross entropy as follows,

$$\mathcal{L}_{cls}^G = \mathbb{E}_{(y_j, y_j^{y_j \neq v_k}, \mathbf{g}_j^{y_j \neq v_k}) \sim (\mathcal{Y}_j, \mathcal{Y}_j', \mathcal{G}_j')} [w * \ell(y_j^{y_j \neq v_k}, \mathbf{g}_j^{y_j \neq v_k})], \tag{4}$$

where $\ell(a, b)$ represents the function of the binary cross entropy defined as $-a * log D_j^{cls}(b) - (1 - a) * log(1 - D_j^{cls}(b))$. $D_j^{cls}(.)$ judges whether the targeted attribute y_j' is edited in $\mathbf{g}_j'$.

Reconstruction loss avoids irrelevant editing and noise in the edited image. $\mathbf{g}_j$ is the result of $g_j^{y_j \neq v_k}$ edited by G_j according to original attribute value y_j, *i.e.* $\mathbf{g}_j = \mathbf{g}_j^{y_j \neq v_k} + G_j(\mathbf{g}_j^{y_j \neq v_k}, y_j)$. The loss is defined as:

$$\mathcal{L}_{rec}^G = \mathbb{E}_{(y_j, \mathbf{x}, \mathbf{g}_j) \sim (\mathcal{Y}_j, \mathcal{X}, \mathcal{G}_j)} [w * \|\mathbf{x} - \mathbf{g}_j\|_1], \tag{5}$$

where $\|\cdot\|_1$ represents L1-norm.

Regularization loss is an additional component considered in our model comparing with StarGAN. As a residual image rather than a whole image is considered in our G_j, a residual image describes the regional transformation on the target attribute. The study (Shen et al., 2017) indicates a residual image should be sparse because it only refers to a small number of effective pixels. Therefore, the regularization loss is defined as:

$$\mathcal{L}_{l1}^G = \mathbb{E}_{(y_j, \mathbf{r}_j^{y_j \neq v_k}) \sim (\mathcal{Y}_j, \mathcal{R}_j')} [w * \|\mathbf{r}_j^{y_j \neq v_k}\|_1] \tag{6}$$

where $\|\cdot\|_1$ is L-1 norm.

3.2.3. Discriminator training

Adversarial and classification loss are considered in the objective function of D_j, defined as:

$$\min_{D_j} \mathcal{L}^D = \mathcal{L}_{adv}^D + \lambda_{cls}^D * \mathcal{L}_{cls}^D \tag{7}$$

where λ_{cls}^D indicate importance of classification losses.

Adversarial Loss for D_j distinguishes the real and generated images. Similar to the adversarial loss used in the generator, we also

quantify the realistic level of images by the weighted logarithm of D_j^{real}. The loss is defined as:

$$\mathcal{L}_{adv}^{D} = \mathbb{E}_{(y_j,\mathbf{x})\sim(\mathcal{Y}_j,\mathcal{X})}[w * log(1 - D_j^{real}(\mathbf{x}))] + \mathbb{E}_{(y_j,\mathbf{g}_j^{y_j\neq v_k})\sim(\mathcal{Y}_j,\mathcal{G}_j')}[w * log(D_j^{real}(\mathbf{g}_j^{y_j\neq v_k}))], \quad (8)$$

Classification Loss for D_j quantifies the correlation between the jth predicted attribute value of input image and the jth real attribute label. The loss is defined as follows:

$$\mathcal{L}_{cls}^{D} = \mathbb{E}_{(y_j,\mathbf{x})\sim(\mathcal{Y}_j,\mathcal{X})}[w * \ell(y_j,\mathbf{x})], \quad (9)$$

4. Experiment

In this section, we empirically evaluate and compare the effectiveness of PESGAN to other related methods on manipulating single and multiple attributes.

4.1. Experimental settings

CelebA contains 202,2599 face images of celebrities, labeled with 40 binary attributes. Eight attributes are chosen for comprehensive comparison in our experiment, including 'Mouth_Slightly_Open' (MSO), 'Bags_Under_Eyes' (BUE), 'Bushy_Eyebrows' (BE), 'Rosy_Cheeks' (RC), 'Eyeglasses' (EG), 'Sideburns' (SB), 'Pale_Skin' (PS), 'Mustache' (MT). The imbalance ratios, defined as the ratio between the number of samples with the majority attribute value and samples with the minority attribute value, of the mentioned attributes are: 1.1, 3.9, 6.0, 14.2, 14.4, 16.7, 22.3 and 23.1 respectively. Both single and multiple attributes manipulation are carried. On the other hand, in order to illustrate the generalization of the GAN methods on a dataset without imbalance attribute value, Radboud Faces Database (RaFD) (Langner et al., 2010) is also considered. RaFD contains 4824 images collected from 67 participants in three different gaze directions and three different angles, displaying eight expressions, including 'Angry', 'Contemptuous', 'Disgusted', 'Fearful', 'Happy', 'Neutral', 'Sad', 'Surprised'. All attribute values are balanced. As the size of samples with different expressions are similar, RaFD is a balanced dataset. The same image preprocessing and dataset partitioning as Choi et al. (2018) are used.

The representative methods of cGAN-based and rGAN-based multiple facial attributes manipulation methods, including StarGAN, AttGAN and IcGAN, are used to compare with our method. All hyper-parameters of StarGAN and AttGAN are set by default. However, the hyper-parameters of IcGAN are fine-tuned since the defaulted settings are determined for a 64 × 64 image, which is different from our experiments. The necessities of our proposed components, including stack structure and weighted learning, on single attribute manipulation are evaluated. Two additional models, *i.e. baseline*, and *baseline+stack*, are also considered in the ablation study. *Baseline* contains only a single network with the residual image generation, while the stacked network without weighted learning is used in *baseline+stack*.

The generators used in our model adopts a encoder–decoder structure. The encoder consists of four convolutional layers, while the decoder contains four transposed convolutional layers. The U-Net structure (Ronneberger et al., 2015) in which the symmetric skip connections are constructed between encoder and decoder is used. On the other hand, four convolutional layers are used by the discriminator. Both D_j^{adv} and D_j^{cls} contains two fully connected layers. Similar to StarGAN and AttGAN, WGAN-GP (Gulrajani et al., 2017) is applied to stabilize the training process of our method. Adam optimizer (Kingma et al., 2015) is used for both generator and discriminator with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ in (2) and (7). We set $\lambda_{cls}^{G} = \lambda_{rec}^{G} = \lambda_{l1}^{G} = \lambda_{cls}^{D} = 10$. The learning rate is set as 2×10^{-5} and decreases by 0.97 every 3000 steps. All methods are updated 100,000 and 50,000 steps for CelebA and RaFD respectively.

Besides the qualitative discussion on generated images, Fréchet Inception Distance score (FID) (Heusel et al., 2017) is applied to quantify the fidelity and diversity of edited images. In addition, the satisfaction of the target attribute modification is quantified by a classifier. An individual ResNet-18 network (He et al., 2016) is trained for each CelebA and RaFD to determine which target attribute values existed in an image.

4.2. Manipulation on single attribute on CelebA

This section discusses the performance on manipulating on a single attribute by our proposed PESGAN, StarGAN, AttGAN and IcGAN. All models are trained as a multiple attributes generator for the eight selected attributes in CelebA. The quality of the generated images is first discussed, and then the images are evaluated by FID and the classification accuracy of the target attribute existed in the generated images.

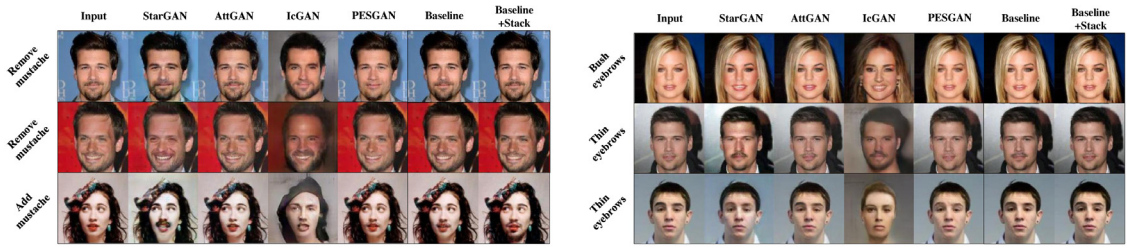
A column in Fig. 6 shows the input and its generated images, while a row represents a selected sample. Obviously, the performance of IcGAN is significantly worse than the cGAN-based models. In most cases, a person in the face images generated by IcGAN is different from the original one. Information may lose during compressing an image to a 100-dimensional space in the encoding process of IcGAN. It may explain why the image reconstruction of IcGAN is not satisfying. StarGAN manipulates the attributes successfully but the facial identities are also changed slightly. The performance of AttGAN is more satisfying comparing with StarGAN. It is mainly because the skip connection in Unet of AttGAN enable the attribute-irrelevant regions of an image to directly pass the manipulation, which preserves the quality of the original image. On the other hand, PESGAN achieves the best performance among all models in terms of the preservation on facial identity and ability of attribute manipulation. The residual image generation used in PESGAN avoids the unnecessary change on the attribute-irrelevant regions as well as the facial identity. Moreover, the difficulty of multi-attribute manipulation is reduced by the divide and conquer method in our method.

The experimental results confirm the three baseline methods suffer from attribute entanglement due to insufficient samples on some attribute values. Other attributes besides target one have been also changed by three baseline models in most generation. For example, the eyebags have been enhanced or weaken in the mustache manipulation of StarGAN, AttGAN, IcGAN shown in Fig. 6(a). In addition, the skin color of the generated images is also different in Figs. 6(a) and 6(b). Among these three baseline methods, IcGAN changes undesired attributes most obviously, which may be because of the information loss caused by the encoding process. By contrast, PESGAN only focus on the target attribute in manipulation and achieve the best results. It is because the deployment of editing each attribute independently avoids *attribute entanglementattribute entanglement* effectively.'

Attribute-irrelevant regions are interfered by baseline models leading to *imprecise editing*. Obvious noise can be observed in the forehead, region under eyes and profile of the persons generated by StarGAN, AttGAN and IcGAN shown in Fig. 6(a). Moreover, the color of the hair and mouth generated by the baseline models are also changed, illustrated in Fig. 6(b). However, the editing of PESGAN concentrates on the target attribute since a residual image is considered.

The performance on minority attribute values of all baseline models is significantly worse than the one on majority attribute values. Fig. 6(a) indicates that the manipulation on mustache is not obvious by all baseline models, but our proposed model still achieves satisfying results. It confirms that the knowledge of minority attribute values can be utilized more effectively in PESGAN (see Fig. 3).

Table 1 shows the FID scores of single attribute manipulation on CelebA. + (−) represents the majority (minority) value for an attribute. FID scores of our proposed PESGAN are significantly lower than other methods in all cases, while IcGAN generally achieves the highest FID score. FID of PESGAN for MSO, BUE, BE and MT are 5.1 (5.4), 5.0 (6.5), 5.2 (8.5) and 5.6 (7.8) for majority (minority) values respectively.



(a) Mustache (b) Bushy_Eyebrows

Fig. 3. Examples of an image generated by StarGAN, AttGAN, IcGAN, and PESGAN for single attribute manipulation on CelebA.

Table 1
FID of single attribute manipulation of StarGAN, AttGAN, IcGAN, and PESGAN on CelebA.

FID	MSO		BUE		BE		RC		EG		SB		PS		MT		Average		
	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−	Total
StarGAN(Choi et al., 2018)	55.0	54.4	63.4	55.7	59.0	53.8	73.1	59.7	85.9	77.8	80.2	60.6	86.5	58.2	83.3	65.8	73.3	60.8	67.1
AttGAN(He et al., 2019)	32.9	33.1	35.1	33.5	36.3	31.2	42.0	31.6	49.4	46.4	45.8	38.8	43.1	31.3	44.3	39.7	41.1	35.7	38.4
IcGAN(Perarnau et al., 2016)	113.5	109.2	118.4	112.4	120.5	113.5	140.0	120.2	146.8	135.0	126.0	115.8	116.9	95.9	126.7	125.7	126.1	116.0	121.1
PESGAN	**5.1**	**5.4**	**5.0**	**6.5**	**5.2**	**8.5**	**9.9**	20.3	**33.9**	**35.8**	**14.3**	**16.6**	**16.7**	21.3	**5.6**	**7.8**	**12.0**	**15.3**	**13.7**
Baseline	15.1	14.5	13.5	12.1	17.2	15.9	30.3	24.1	47.4	43.5	40.8	32.8	29.4	21.3	35.6	29.2	28.7	24.2	26.5
Baseline + stack	8.2	7.6	9.4	7.5	12.5	8.5	27.2	**17.1**	42.3	37.1	29.1	20.7	22.9	**13.5**	25.7	19.7	22.2	16.5	19.4

* +(−) represents the attributes with the majority (minority) value.

These scores are dramatically lower than the ones achieved by other methods. The performance of three baseline methods is significantly better in manipulating majority attribute values than minority values, *i.e.* the FID scores of + is larger than the ones of −. By contrast, the performance of our method is more balanced on both majority and minority values, and manipulates the minority attribute value slightly better. This results suggest that our method successfully reduce the influence of minority attribute values to training. The results confirms our model outperforms other baseline models in terms of the fidelity and diversity of generated images. It is because our divide-and-conquer method and residual image generation ease the difficulty of editing each attributes. Moreover, the weighted learning is used to improve the performance on editing minority attribute values.

An ablation study, including *baseline* and *baseline+stack*, is carried out to illustrate the importance of components of our model. The FID scores shown in Table 1 indicates that the quality of the images generated by PESGAN is higher than *baseline* and *baseline+stack*. *Baseline* performs the worst due to the ability limitation of a single network. *Baseline+stack* performs better than *baseline* in terms of not changing the undesired attributes such as eyebrows, eyebags and mustache since of its stacked structure, shown in Fig. 6. However, *baseline+stack* cannot remove mustache and thin eyebrows as *baseline*.

4.3. Manipulation on multiple attributes on CelebA

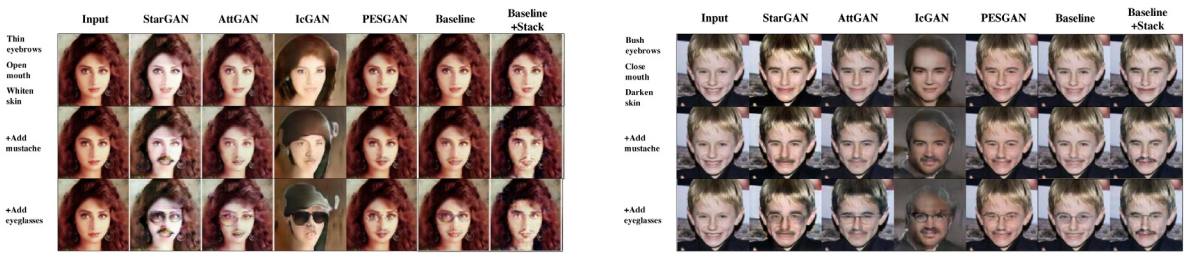
The imbalance problem is exacerbated when multiple attributes are considered since the size difference between the samples containing majority attribute values and the ones with minority attribute values increases. Samples containing combination of minority attribute values may be insufficient. This experiment focus on the manipulation on multiple attributes, which are BE, MSO, PS, MT, and EG. Three combination on these five attributes are considered: BE+MSO+PS, BE+MSO+PS+MT, and BE+MSO+PS+MT+EG. The imbalance ratios of the combinations are 252.9, 11,494.4, and 37,552.0 respectively. These ratios increase dramatically when more attributes are considered. Similar to the previous experiment, all models are trained to manipulate the eight selected attributes.

The experimental results of our model and baseline models are illustrated in Fig. 7. The first five columns in Fig. 7 show the input and generated images by four models, while a row represents the results of target attributes. Generally, the quality of images generated by all methods with multiple attributes manipulation is worse significantly than the single one. The quality decreases with the increase of edited attribute number, illustrated by Figs. 4(a) and 4(b). However, the image quality of our method is similar to the other methods although the errors may be accumulated in the cascade framework. It may be because that our model work well on single-attribute editing tasks as they are relatively simpler than the multi-attribute one. The results also confirm that generating a residual image reduces the loss in image reconstruction.

The results show that StarGAN, AttGAN, IcGAN not only change the target attributes but also other attributes. For example, in Fig. 4(b), all three baseline models weaken the attribute eyebags. This indicates the problem of *attribute entanglement* of the baseline methods. In addition, the attribute-irrelevant region are also interfered in three baseline methods. Fig. 4(a) are examples which indicate obvious noise can be found in the forehead. Furthermore, the baseline methods change the color of mouth, hair and even background in Fig. 4(a). Moreover, the performance of StarGAN, AttGAN, and IcGAN on minority attributes values is also not satisfying. As shown in Figs. 4(a) and 4(b), darken skin is not edited obviously by using the baseline methods. On the other hand, the interfere of the increase of target attribute number on our proposed method is much smaller comparing with other methods. The quality on manipulating attributes as well as attribute-irrelevant information (*i.e.* facial identity and background) is satisfying. Moreover, PESGAN works well on the minority attribute values.

Table 2 shows the FID scores of multiple attributes manipulation on CelebA. The FID scores of PESGAN is the smallest among all methods in all situations. Moreover, FID scores of our method increase more stably with the increase of edited attribute number than other three base methods. In specific, IcGAN achieves an dramatically high FID score. The experimental results confirms that PESGAN performs well on the multiple attributes manipulation.

The similar problems of the existing models can be observed in the results of *baseline* and *baseline+stack*, shown in the final two columns of Fig. 7. The problem of attribute entanglement occurs in *baseline* and its results are interfered by undesired attributes, *e.g.*, the attribute eyebags is weaken in Fig. 4(b). Moreover, the performance on majority attribute values of both *baseline* and *baseline+stack* is better than the one on minority attribute values due to the domination of the majority attribute values in training. Fig. 4(a) illustrates that *baseline+stack* hardly



(a) (b)

Fig. 4. Examples of an image generated by StarGAN, AttGAN, IcGAN, and PESGAN for multiple attributes manipulation on CelebA . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

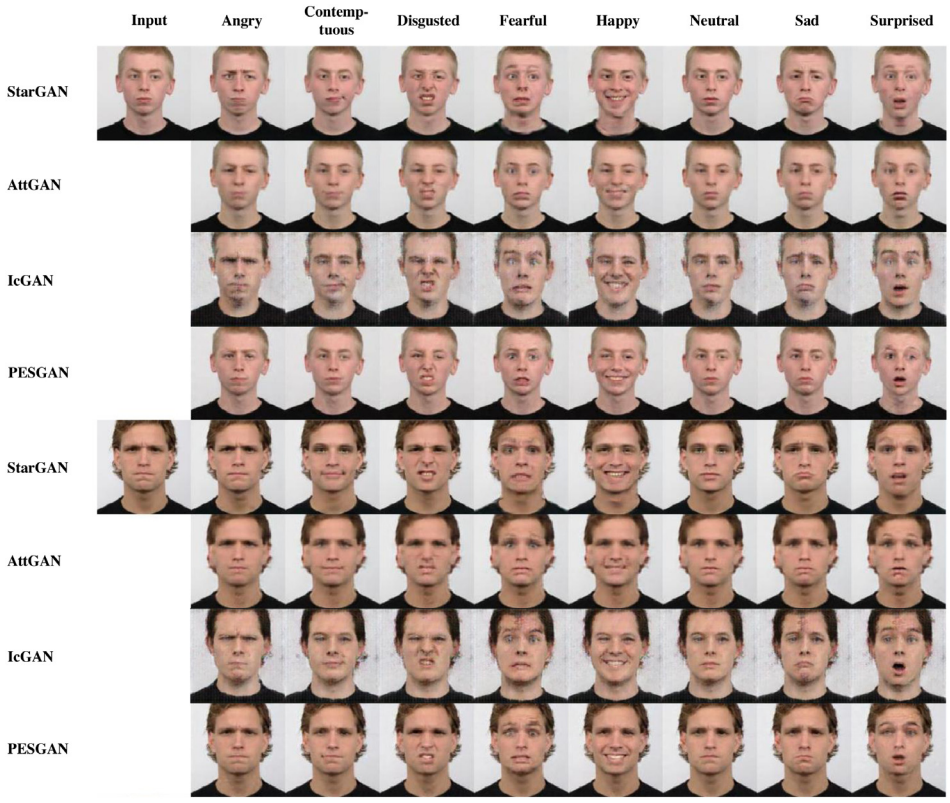


Fig. 5. Examples of an image generated by StarGAN, AttGAN, IcGAN, and PESGAN for single attributes manipulation on RaFD.

Table 2
FID of multiple attributes manipulation of StarGAN, AttGAN, IcGAN, and PESGAN on CelebA.

FID	BE+MSO+PS	BE+MSO+PS+MT	BE+MSO+PS+MT+EG	Average
StarGAN(Choi et al., 2018)	80.8	108.4	138.7	109.3
AttGAN(He et al., 2019)	63.9	97.6	140.9	100.8
IcGAN(Perarnau et al., 2016)	160.7	205.2	278.5	214.8
PESGAN	**28.6**	**37.4**	**56.8**	**40.9**
Baseline	55.7	71.9	102.4	76.7
Baseline + stack	36.1	52.3	84.1	57.5

darkens the skin. PESGAN performs well on editing both majority and minority attribute values.

4.4. Manipulation of single attribute on RaFD

The experiments are also carried out on RaFD in order to illustrate the generation ability of our proposed method to a balance dataset. Each generation method is trained to edit eight expressions. Four examples of the results are displayed in Fig. 5. Columns show the input and generated images with eight different expressions, while a row represents the results produced by a model.

The experimental results indicate that StarGAN and AttGAN as well as our method work well when the attribute values are balanced and only single attribute manipulation is considered. PESGAN is not superior in this situation. On the other hand, the performance of IcGAN is still not good. Its generated images are still in low quality. Similar results can be obtained from FID, shown in Table 3. The average FID scores of StarGAN, AttGAN and PESGAN are around 20, but IcGAN is nearly 40.

Table 4 presents the classification accuracy on RaFD. The accuracy of PESGAN and StarGAN is similar, which is 56.5 and 57.3. IcGAN

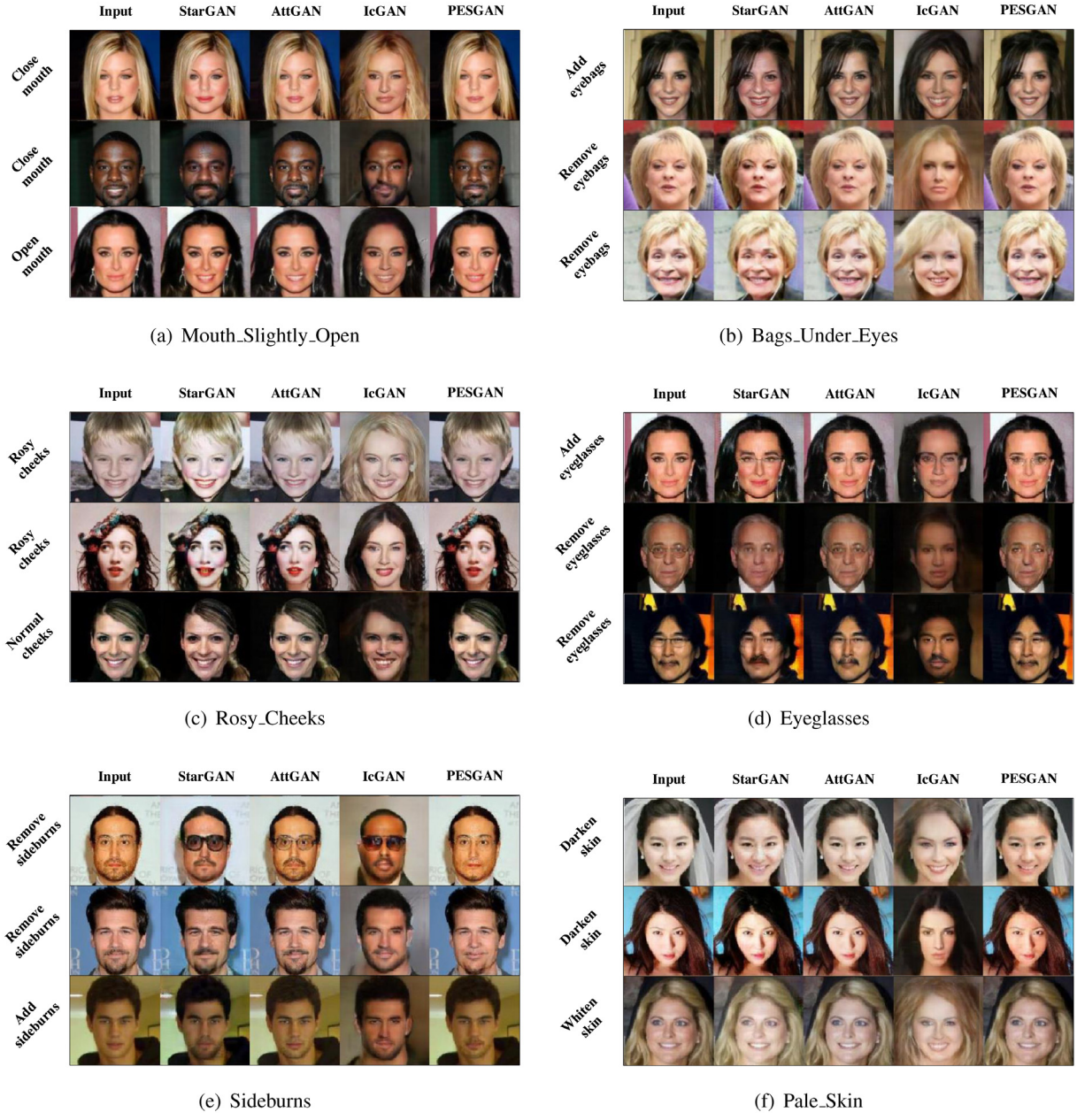


Fig. 6. Examples of an image generated by StarGAN, AttGAN, IcGAN, and PESGAN for single attribute manipulation on CelebA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
FID of single attribute manipulation of StarGAN, AttGAN, IcGAN, and PESGAN on RaFD.

FID	Angry	Contemptuous	Disgusted	Fearful	Happy	Neutral	Sad	Surprised	Average
StarGAN(Choi et al., 2018)	25.4	**21.1**	17.1	15.3	**18.3**	**14.5**	**15.6**	27.8	**19.4**
AttGAN(He et al., 2019)	**23.4**	22.7	18.4	16.8	22.4	16.2	19.9	29.6	21.2
IcGAN(Perarnau et al., 2016)	41.0	40.5	37.5	37.1	43.5	35.3	38.4	42.2	39.4
PESGAN	24.8	23.6	**15.5**	**14.3**	21.1	15.7	18.2	**24.7**	19.7

has the worst performance again and its accuracy is only 52.3. In conclusion, our proposed model still work well in RaFD. However, the performance of PESGAN, StarGAN and AttGAN is not significantly different on single attribute manipulation training from balanced attribute values.

4.5. Training cost

The cost of training required by the all methods is evaluated in GTX 1080. The training time (in hour) and the number of parameters used in the models on editing eight attributes for CelebA dataset are showed in Table 5. Although our proposed PESGAN requires eight generators, its training time is 29 h, which is shorter than AttGAN (30 h) and IcGAN (38 h). It confirms our expectation that the total training time is shorter even though a number of base models are used since the editing problem of a base model is significantly simpler than the original problem. Similarly, the parameters required by our method are more than StarGAN which deploys many residual blocks in the network in order to reduce the number of parameters. As a network

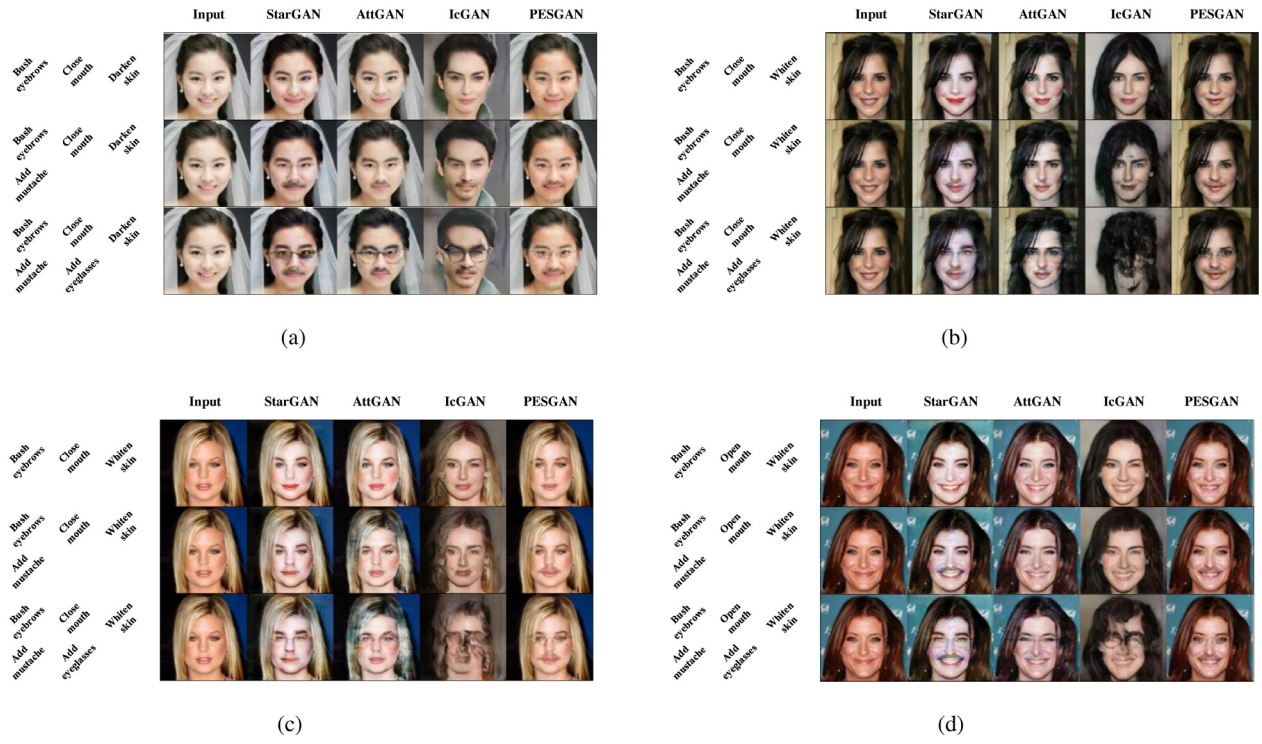


Fig. 7. Examples of an image generated by StarGAN, AttGAN, IcGAN, and PESGAN for multiple attributes manipulation on CelebA.

Table 4
Classification accuracy (%) of the target attribute existed in the images generated by StarGAN, AttGAN, IcGAN, and PESGAN for single attribute manipulation on RaFD.

Accuracy	Angry	Contemptuous	Disgusted	Fearful	Happy	Neutral	Sad	Surprised	Average
StarGAN(Choi et al., 2018)	62.5	**55.4**	64.9	69.8	**70.2**	**25.6**	**65.0**	45.2	**57.3**
AttGAN(He et al., 2019)	**67.3**	52.9	62.3	63.0	60.4	22.6	55.3	41.1	53.1
IcGAN(Perarnau et al., 2016)	63.2	53.8	62.7	59.5	62.9	21.2	56.1	39.0	52.3
PESGAN	63.6	48.0	**68.9**	**70.7**	65.3	24.5	59.5	**51.2**	56.5

Table 5
Parameter number and training time of StarGAN, AttGAN, IcGAN, and PESGAN for eight-attribute manipulation on CelebA.

Model	Parameter number	Training time
StarGAN(Choi et al., 2018)	8.4M	14 h
AttGAN(He et al., 2019)	43.3M	30 h
IcGAN(Perarnau et al., 2016)	130.0M	38 h
PESGAN	24.8M	29 h

with a simple structure is used in each generator in PESGAN, its total parameter number is significantly less than AttGAN and IcGAN, which require a complicated network.

5. Conclusions and future work

This study addresses three major challenges in multiple facial attributes manipulation, which are attribute entanglement, imprecise editing and bad performance on minority attribute values. We propose PESGAN which stacks a number of individual generators. As each generator manipulates only one attribute independently, the imbalance problem on attribute values will not be deteriorated when multiple facial attributes manipulation is considered. Moreover, attribute entanglement is also avoided since every attribute is edited individually and the influence to others is minimized. In addition, the residual image generation is deployed to reduce the interference on attribute-irrelevant region. The superior of our proposed model is confirmed by the experimental results on CelebA and RaFD. The results confirm that our model remarkably outperforms the previous leading methods, StarGAN, AttGAN and IcGAN, in terms of image quality and attribute editing accuracy, especially on multiple attributes manipulation. Moreover, the training cost of PESGAN is competitive to other methods.

Our framework can be considered to manipulate multiple attributes with continuous values. Editing continuous attribute is a tough question as an attribute should be modified as a certain degree. Our divide-and-conquer framework allows that a base generator focuses on only continuous attribute in a multiple attribute editing problem, which reduces the learning difficulty and may increase the quality of generated images.

CRediT authorship contribution statement

Patrick P.K. Chan: Conceptualization, Methodology, Investigation, Writing – original draft, Funding acquisition. **Xiaotian Wang:** Methodology, Software, Investigation, Writing – original draft. **Zhe Lin:** Methodology, Validation, Investigation, Writing – review & editing. **Daniel S. Yeung:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

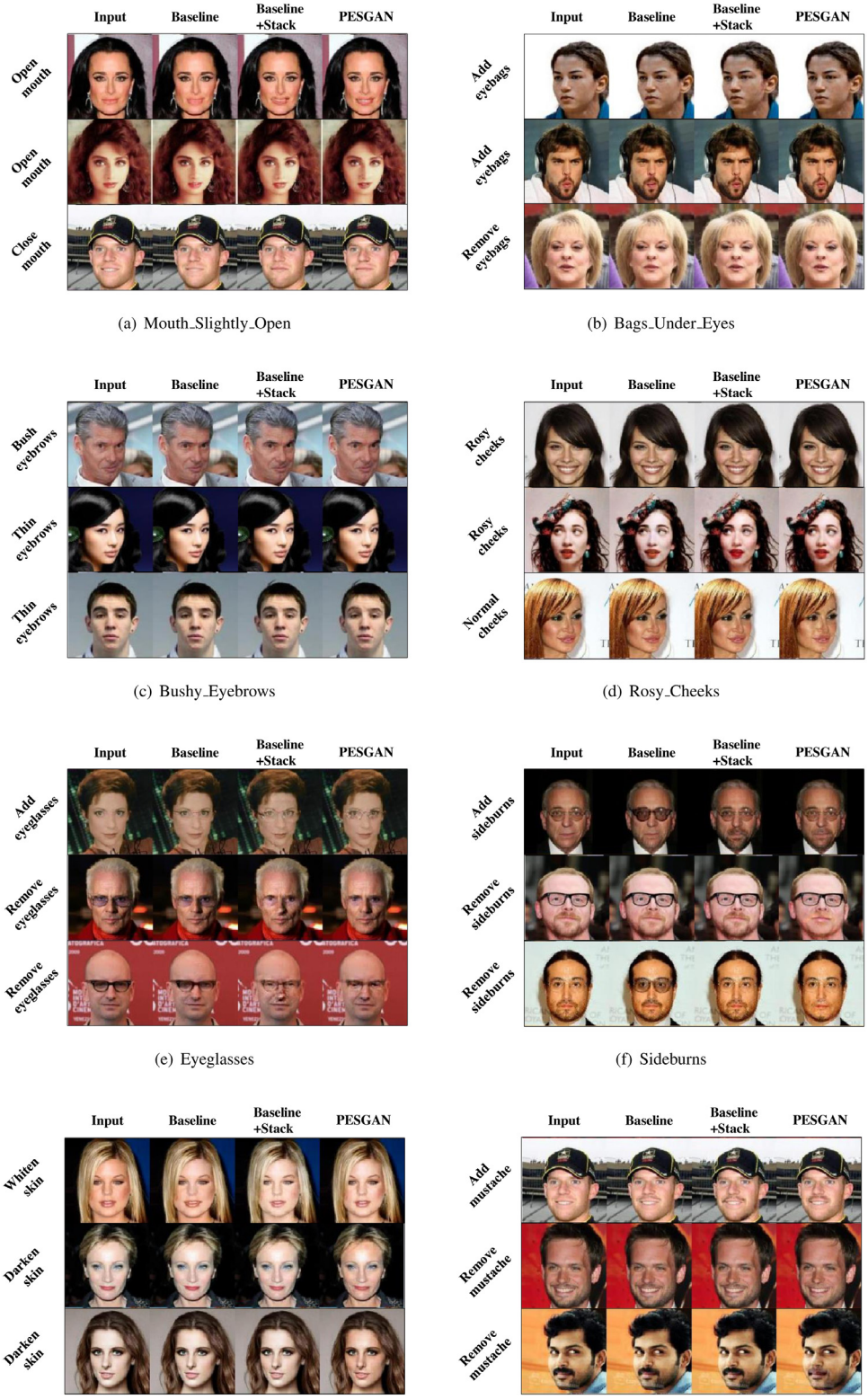
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by Natural Science Foundation of Guangdong Province, China (No. 2018A030313203), and the Fundamental Research Funds for the Central Universities, China (South China University of Technology) (No. 2018ZD32).

Appendix. Additional experimental results

Figs. 6 and 7 show the results generated by three baseline methods and our proposed PESGAN on CelebA. In Fig. 6(e), the eyeglasses of synthesized image are changed improperly by the baseline methods, *i.e.* the generated images are far away from the target of editing sideburns.



(a) Mouth_Slightly_Open (b) Bags_Under_Eyes

(c) Bushy_Eyebrows (d) Rosy_Cheeks

(e) Eyeglasses (f) Sideburns

(g) Pale_Skin (h) Mustache

Fig. 8. Examples of an image generated by baseline, baseline + stack, and PESGAN for single attribute manipulation on CelebA.

Fig. 9. Examples of an image generated by baseline, baseline + stack and PESGAN for multiple attributes manipulation on CelebA.

Fig. 6(f) indicates that the baseline methods are not capable of editing minority attribute class, such as pale skin, obviously. The original color of mouth and background cannot be preserved completely by the baseline methods in Fig. 6(c). By contrast, our PESGAN performs well in nearly all cases. Fig. 7 indicates the similar problems can be found in editing multiple attributes of the baseline methods. However, PESGAN can relieve these problems and shows its superiority.

Figs. 8 and 9 depict the necessities of our proposed components, including stack structure and weighted learning. In general, all models manipulate image precisely since of using residual image generation. However, the baseline has the worst performance generally due to the difficulty of editing various attributes by one single network. Besides, the baseline and baseline + stack perform badly on the minority attribute values. On the contrary, our proposed PESGAN not only edits attributes precisely, but also has a balanced performance on both the majority and minority attribute value because of weighted learning.

References

Choi, Yunjey, et al., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8789–8797.

Emami, Hajar, et al., 2020. Spa-gan: Spatial attention gan for image-to-image translation. IEEE Trans. Multimedia (TMM).

Frid-Adar, Maayan, et al., 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 1312.6114, 321–331.

Goodfellow, Ian, et al., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS).

Gulrajani, I., Ahmed, F., et al., 2017. Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems (NIPS).

He, Kaiming, et al., 2016. Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

He, Zhenliang, et al., 2019. Attgan: Facial attribute editing by only changing what you want. IEEE Trans. Image Process. (TIP) 28, 5464–5478.

Heusel, Martin, et al., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NIPS).

Huang, Xun, et al., 2017. Stacked generative adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Isola, Phillip, et al., 2017. Image-to-image translation with conditional adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kingma, D.P., et al., 2015. Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR).

Langner, O., Dotsch, R., et al., 2010. Presentation and validation of the radboud faces database. Cogn. Emot. 1377–1388.

Li, Mu, Zuo, Wangmeng, et al., 2016. Deep identity-aware transfer of facial attributes. arXiv preprint 1610.05586.

Li, Xiaoqiang, et al., 2018a. SCGAN: Disentangled representation learning by adding similarity constraint on generative adversarial nets. IEEE Access 147928–147938.

Li, Zejian, et al., 2018b. Unsupervised disentangled representation learning with analogical relations. In: The International Joint Conference on Artificial Intelligence (IJCAN). pp. 2418–2424.

Mirza, Mehdi, et al., 2016. Conditional generative adversarial nets. arXiv preprint 1411.1784.

Perarnau, Guim, et al., 2016. Invertible conditional gans for image editing. In: Advances in Neural Information Processing Systems (NIPS).

Pfeiffer, Micha, et al., 2019. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In: The International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Ronneberger, Olaf, et al., 2015. U-net: Convolutional networks for biomedical image segmentation. In: The International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241.

Shen, Wei, et al., 2017. Learning residual images for face attribute manipulation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1225–1233.

Tran, Luan, Yin, X., et al., 2017. Disentangled representation learning GAN for pose-invariant face recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Wang, W., Cui, Z., et al., 2016. Recurrent face aging. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Xu, T., Zhang, et al., 2018. Attngan: fine-grained text to image generation with attentional generative adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D., 2017. Stack-GAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV).

Zhang, Gang, et al., 2018a. Exprgan: Facial expression editing with controllable expression intensity. In: The Association for the Advance of Artiicial Intelligence (AAAI).

Zhang, Gang, et al., 2018b. Generative adversarial network with spatial attention for face attribute editing. In: The European Conference on Computer Vision (ECCV).

Zhu, Jun-Yan, et al., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: The IEEE International Conference on Computer Vision (ICCV). pp. 2223–2232.