Neural Networks 133 (2021) 220–228

Contents lists available at ScienceDirect

Neural Networks

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

2020 Special Issue

ClsGAN: Selective Attribute Editing Model based on Classification Adversarial Network

Ying Liu [a,b], Heng Fan [c], Fuchuan Ni [a,b], Jinhai Xiang [a,b,*]

[a] College of Informatics, Huazhong Agricultural University, Wuhan, 430070, China
[b] Hubei Engineering Technology Research Center of Agricultural Big Data, Huazhong Agricultural University, Wuhan, 430070, China
[c] Department of Computer Science, Stony Brook University, Stony Brook, 11794, USA

ARTICLE INFO

Article history:
Available online 10 November 2020

Keywords:
GAN
Attribute editing
ClsGAN
Upper convolution residual network (Tr-resnet)
Attribute adversarial classifier (Atta-cls)

ABSTRACT

Attribution editing has achieved remarkable progress in recent years owing to the encoder–decoder structure and generative adversarial network (GAN). However, it remains challenging to generate high-quality images with accurate attribute transformation. Attacking these problems, the work proposes a novel selective attribute editing model based on classification adversarial network (referred to as ClsGAN) that shows good balance between attribute transfer accuracy and photo-realistic images. Considering that the editing images are prone to be affected by original attribute due to skip-connection in encoder–decoder structure, an upper convolution residual network (referred to as Tr-resnet) is presented to selectively extract information from the source image and target label. In addition, to further improve the transfer accuracy of generated images, an attribute adversarial classifier (referred to as Atta-cls) is introduced to guide the generator from the perspective of attribute through learning the defects of attribute transfer images. Experimental results on CelebA demonstrate that our ClsGAN performs favorably against state-of-the-art approaches in image quality and transfer accuracy. Moreover, ablation studies are also designed to verify the great performance of Tr-resnet and Atta-cls.

1. Introduction

Attribute editing (also termed as attribute transfer) aims to change one or more attributes of images (e.g., hair color, sex, style, etc.) while other attributes remain. The key of attribute editing is to achieve high quality and accurate attribute transfer of generated images. In recent years, generative adversarial network (GAN) (Goodfellow et al., 2014) has greatly advanced the development of attribute editing. Inspired by this, numerous approaches (Choi et al., 2018; He et al., 2017; Li et al., 2019, 2016; Liu et al., 2019; Xie et al., 2019; Zhou et al., 2017) have been proposed to change local (e.g., hair color, adding accessories, altering facial expressions, etc.) or global (e.g., gender, age, style, etc.) attributes of images.

In addition, in order to obtain accurate attribute transfer images, encoder–decoder architectures Hinton and Zemel (1994a) have been used in attribute editing. Despite promising performance, the method may result in poor quality of generated image because of the bottleneck layer. To address the issue, skip-connection is applied to encoder–decoder architecture for high-quality image (He et al., 2017; Liu et al., 2019). Nevertheless, the use of skip-connection brings about the trade-off between image quality and accuracy (Liu et al., 2019), i.e., it generates high-quality images at the cost of low attribute accuracy.

Through an in-depth empirical investigation of GAN model (Goodfellow et al., 2014), in addition to the original image, the generated image is also required to be fed to the discriminator for learning the defects of generated images during the training of discriminator. This way, the discriminator is able to guide the optimization of generator according to the defect information. Besides, for attribute classifier in attribute editing, most recent approaches (Choi et al., 2018; He et al., 2017; Li et al., 2019; Liu et al., 2019; Xie et al., 2019) only take as input the original images. Nevertheless, these methods ignore the positive role of generated images on enforcing attribute transfer accuracy when training the classifier.

In order to address the aforementioned issues, a novel selective attribute editing model based on classification adversarial network (referred to ClsGAN for short) is proposed. The key innovation of ClsGAN is an attribute adversarial classifier (referred to Atta-cls for short) that aims at enhancing the classification performance. Sharing similar spirit of GAN model (Goodfellow et al., 2014), Atta-cls is implemented as an adversarial network of

* Corresponding author at: College of Informatics, Huazhong Agricultural University, Wuhan, 430070, China.
E-mail address: jimmy_xiang@mail.hzau.edu.cn (J. Xiang).

https://doi.org/10.1016/j.neunet.2020.10.019
0893-6080/

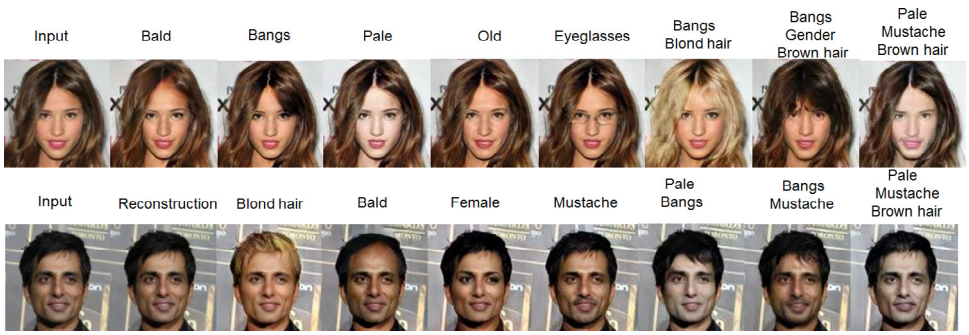


Fig. 1. Illustration of generated images with the proposed ClsGAN. These generated images demonstrate high quality and accurate attribute transfer from the visual perspective. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

image attributes. During the training of Atta-cls, both the original and generated images are fed to the classifier in which it specifies that the attributes of generated image are indistinguishable (similar to the fake property in GAN).

In addition, a detailed empirical analysis on limitation of skip-connect in deep encoder–decoder structure (Hinton & Zemel, 1994a) is conducted, and the main reason is caused by transmitting both source attribute and details of images into the decoder. Since more information of source images is encoded, it may decrease the target attribute information, resulting in degraded performance. Motivated by the residual neural network (Xie et al., 2017), an effective upper convolution residual network (referred to Tr-resnet for short) which is as a decoder is proposed to get more the target attribute information. Tr-resnet is able to selectively acquire source image and target label information by combining both input and output of the upper convolution residual blocks, leading to high-quality image generation with accurate attribute editing.

Moreover, drawing on the practices of Radford et al. (2016) and Xie et al. (2019), ClsGAN takes the images as input into two separate encoders (attribute encoder and content encoder) to decouple entanglement between the attribute and unchanged content information. To keep labels continuous, encoded attribute information (i.e., the output of attribute encoder) is also employed to approximate to reference labels. As shown in Fig. 1, the proposed ClsGAN generates photo-realistic images with accurate transfer attributes visually.

In summary, the contributions of this work are three-fold:

- A novel ClsGAN is proposed, which demonstrates significant improvement in realistic image generation with accurate attribute transfer. In particular, the method presents a simple, yet effective upper convolution residual network (Tr-resnet) to alleviate the limitation of skip-connection in encoder–decoder structure.
- In order to improve attribute transfer accuracy, an attribute adversarial classifier (Atta-cls) is developed to guide the generator by learning defects of attribute transfer images.
- Extensive quantitative and qualitative experimental results in face attribute editing demonstrate that the proposed ClsGAN outperforms other state-of-the-art approaches. Furthermore, it is also directly applicable for style manipulation.

The rest of this paper is organized as follows: Section 2 discusses the related work of this paper. Section 3 illustrates the proposed approach in detail. Experimental results are demonstrated in Section 4, followed by the conclusion in Section 5.

2. Related works

Generative adversarial networks (GANs) (Goodfellow et al., 2014), a special case of Artificial Curiosity (Schmidhuber, 2020), are defined as a minimax game with a generator and a discriminator in which the generator generates images as photo-realistic as possible and the discriminator tries to distinguish the synthetic images from the original images. Since then, various GANs and GAN-like variants are proposed to enforce the quality of image or stability of training, including designing novel generator/discriminator architectures (Karras et al., 2019a; Larsen et al., 2016; Radford et al., 2016), the choice of loss functions (Nowozin et al., 2016), the study of regularization techniques (Arjovsky et al., 2017; Gulrajani et al., 2017; Miyato et al., 2018). What is more, Hinton and Zemel (1994b) and Kingma and Welling (2013) introduce an encoder–decoder structure to obtain images' higher-level semantic information and render reconstructed images. VAE/GAN (Larsen et al., 2016) combines VAE (Kingma & Welling, 2013) with GAN (Goodfellow et al., 2014) to modify latent expressions of images by reconstructing loss and adversarial loss. Progressive GAN (Karras et al., 2018) and StyleGAN (Karras et al., 2019b) adopt progressive growth method through layer-by-layer to achieve both style transfer and high-resolution images generation. GANs have been applied to various fields of the computer vision, e.g., image generation (Brock et al., 2018; Goodfellow et al., 2014; Radford et al., 2016), image transfer (Isola et al., 2017; Zhu et al., 2017), super-resolution image (Ledig et al., 2017), image deblurring (Kupyn et al., 2018).

Meanwhile, CGAN (Mirza & Osindero, 2014) takes the reference label as inputs of generator and discriminator to produce specific images that are consistent with the label. Inspired by CGAN, the community makes a large number of contributions in style transfer (Isola et al., 2017; Zhu et al., 2017) and attribute editing (Li et al., 2016; Zhou et al., 2017). About style transfer, Pix2pix (Isola et al., 2017) and CycleGAN (Zhu et al., 2017) realize mutual transformation between two domains about paired and unpaired data respectively. There are also some double domains' transformation models (Li et al., 2016; Zhou et al., 2017) in attributing editing. However, the number of models increases exponentially with the increase of domains by double domains' transformation method, which is not universal and leads to model overfitting and poor generalization ability.

To address the issue, recent methods mostly employed a classifier to realize attribute classification and transformation. StarGAN (Choi et al., 2018) takes domain classification restriction to control the attribute transformation of images, along with reconstruction loss, adversarial loss and classification loss. Notably, He et al. (2017), Liu et al. (2019) and Ronneberger et al. (2015) apply the skip-connection or its variants with the encoder–decoder structure to render photo-realistic images. To avoid the effects of irrelevant attributes, on the one hand, STGAN (Liu et al., 2019) and RelGAN (Wu et al., 2019) both take difference attribute labels as the input. On the other, AME-GAN (Xie et al., 2019) and AGUIT (Li et al., 2019) both separate the input images into image attribute part and image background part on manifolds to

avoid entanglement. StyleGAN (Karras et al., 2019b) introduces a progressive growth method and achieves excellent performance. However, this approach is inefficient due to heavy computation burden. In addition, this paper mainly focuses on learning the styles from the other image that is different from the goal of our paper. Considering these reasons, we do not include it for comparison. In this work, we present ClsGAN which applies Tr-resnet and attribute adversarial classifier to improve image quality and attribute transfer accuracy.

3. The proposed method

In this section, ClsGAN for arbitrary attribute editing is described in detail. Section 3.1 introduces the proposed upper convolution residual network (Tr-resnet). Section 3.2 illustrates attribute continuity processing. After that, an attribute adversarial classifier (Att-cls) is employed to enhance attribute transformation accuracy in Section 3.3. The overall network structure and loss function of our ClsGAN are presented in Sections 3.4 and 3.5.

3.1. Upper convolution residual network (Tr-resnet)

The skip-connection (He et al., 2017) has been proven to be beneficial to improve the quality of generated image. However, this improvement is obtained at the sacrifice of attribute classification performance. In order to solve this problem, the work of He et al. (2017) introduces selective transfer units as a novel skip-connection structure(referred to as STU in short). Nevertheless, the STU requires more parameters and computation resource, which severely limits the application.

In this paper, by empirically investigating skip-connection, the reason causing limitation is that the incorporation of source attribute information weakens the target attribute information in decoder. Motivated by the residual structure in Liu et al. (2019), the upper convolution residual network (Tr-resnet) is proposed. Through using Tr-resnet block as a basic unit, a simple yet effective decoder is developed. As shown on the top right of Fig. 2, each Tr-resnet block takes the combination of the layer's input and output as the unit's output. Furthermore, to effectively make use of resource image and target attribute information, Tr-resnet applies the weighting strategy to resource image information from encoder, input and output information of each Tr-resnet block in a special unit. Mathematically, the Tr-resnet is represented as follows:

$$y^*_{l-1} = \text{Transpose}(y_{l-1}) \tag{1}$$

$$f_l = \begin{cases} \alpha \cdot y^*_{l-1} + (1-\alpha)\cdot y_l & (l \in \{1,2,4,5,6\}) \\ \alpha \cdot y^*_{l-1} + (1-\alpha)\cdot y_l + \beta \cdot x_2 & (l = 3) \end{cases} \tag{2}$$

where y_l denotes the Tr-resnet feature of the lth layer, Transpose $(\cdot)$ represents transposed convolution operation that matches the size between input y_{l-1} and output y_l. x_2 denotes the encoder feature of the 2nd layer, f_l denotes the output of lth Tr-resnet block. In Eq. (2), when $l = 3$, the Tr-resnet block takes the weighted sum of the 2nd layer feature map information of the encoder, the 3rd layer input and output information of Tr-resnet as output. When $l \neq 3$, the output of Tr-resnet block is only the incorporation of information about the input and output of lth layer. The model initializes $\alpha = (a_1, a_2, \ldots, a_s)$, $\beta = (b_1, b_2, ..b_s)$, where $a_i, b_i \sim$ *uniform*$(0, 1)$ and s is the number of feature map in y_l or x_2.

3.2. Attribute continuity processing

The approaches in Choi et al. (2018) and Liu et al. (2019) are able to generate dual-domain (0 or 1) transfer images. However, the methods are difficult to render various images with the same attribute, and attribute continuity cannot be guaranteed. To solve this issue, the work of He et al. (2017) employs a style controller to realize multi-modal transformation for a specific attribute on the basis of the source model, achieving good performance. Motivated by this, the same is utilized to control the attribute continuity in this work. In specific, the attribute value is obtained by approximating encoded attribute label to the reference attribute label. In detail, the optimization object is formulated as follows:

$$L_a = \|l_r - E_a(x_r)\|_1 \tag{3}$$

where x_r and l_r denote source image and reference label, respectively. E_a represents attribute encoder in generator, which is a general convolution neural network and takes Convolution-InstanceNorm-ReLU as a unit. The output of E_a has same size with reference label l_r. $\|\cdot\|_l$ denotes l_1 loss.

3.3. Attribute adversarial classifier (Atta-cls)

For attribute classifier, most existing approaches (Choi et al., 2018; He et al., 2017; Liu et al., 2019) only take the source image as input and then exploit the optimized classifier to improve the generator. However, it is difficult for these approaches to discover the attribute difference between the generated images and source images. Inspired by the GAN model (Goodfellow et al., 2014) that optimizes the generator according to the deficiency of generated image learned from the discriminator, an attribute adversarial classifier (Atta-cls) is proposed based on the adversarial method.

In our model, the attribute classifier is designed as an adversarial network. The source image and the generated image are both fed to optimize the classifier, and then the generator is trained according to the attribute defects of generated image.

Specifically, the target of classifier first evaluates all attribute's distinguishability (similar to the fake/real nature in GAN), and then focuses on the single attributes. When the input is source images, ideally the category is distinguishable (the value is defined as 1 or true) and at the same time the single attribute value should be consistent with label. So the classifier needs to optimize the whole attribute and all single attribute for the source images. In contrast, classifier only needs to assume that it is inseparable for generated images (the value is 0 or false), so the remaining single attributes are considered needlessly. The detailed operation is shown in Fig. 2 (bottom row). Meanwhile, in order to maintain the stability of the model, ClsGAN adds a penalty function for classification loss. The concrete operation is implemented by the loss function. The loss functions of attribute adversarial net (about generator and classifier) are as follows:

$$loss_c = E_{x_r \sim P_{data}}\{t_r^T \log(C(x_r))\} + E_{x_f \sim P_g}\{(1 - t_f^{(1)}) \log(1 - C(x_f)^{(1)})\} \tag{4}$$

$$L_{Cd} = -loss_c + \lambda E_{x^*}[(\|\nabla x^* C(x^*)\|_2 - 1)^2] \tag{5}$$

$$L_{Cg} = E_{x_r \sim P_{data}, G(x_r, l_f) \sim P_g}\{t_f^T \log(C(G(x_r, l_f)))\} \tag{6}$$

where λ is the scale of gradient penalty and sets to 30. L_{Cd} and L_{Cg} denote the loss functions when training classifier(C) and generator about attributes. (x_r, t_r) and (x_f, t_f) correspond to the image and label of source domain P_{data} and generation domain P_g respectively, where $t_r/t_f \in R^{n+1}$. Specially, the first element of t_r/t_f is used to evaluate whether the whole attribute is distinguishable or not, define $t_r^{(1)} = 1, t_f^{(1)} = 0$, and the remaining n-dimensions vector represents image's attributes label (source attribute label l_r or target attribute label l_f). $C(x_f)^{(1)}$ denotes the first element in vector $C(x_f)$, T stands for transpose operation. E_{x^*} denotes gradient penalty term about x^* which is obtained by line sampling between the original and the generated images.

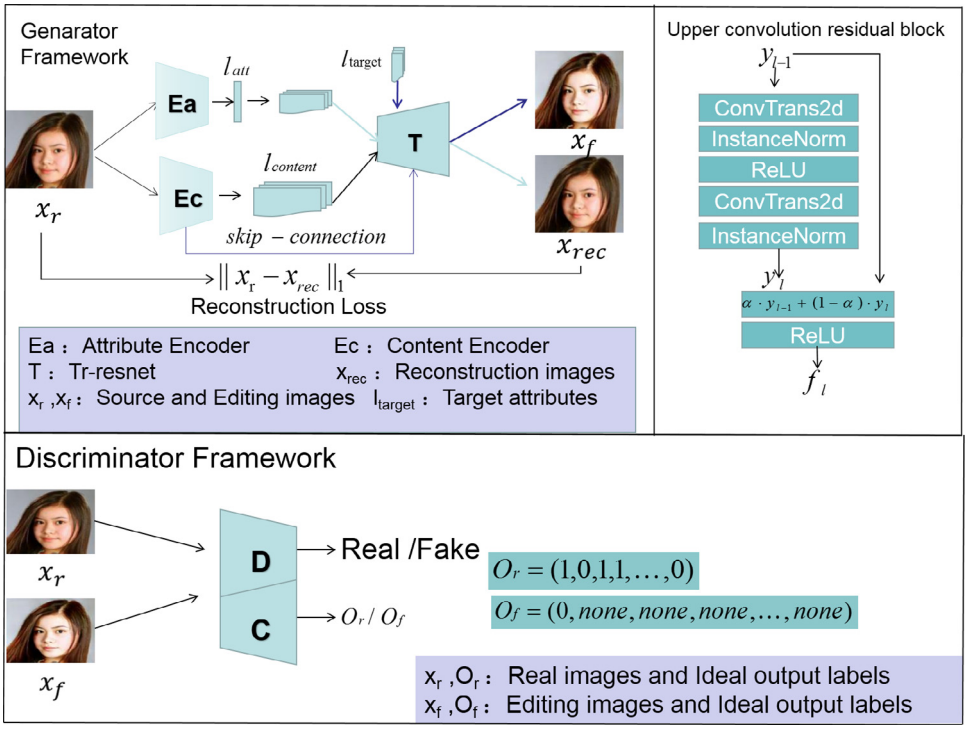


Fig. 2. The structure of ClsGAN, which mainly includes the framework of generator and discriminator. The generator is composed of two encoders and a Tr-resnet(T), which consist of a series of convolution layer and upper convolution residual block (upper right) respectively. Discriminator is composed of classifier C (which is the attribute discriminator of Atta-cls) and adversarial discriminator D, whose parameters are shared.

3.4. Network structure

Fig. 2 shows the framework of ClsGAN, in which the generator is comprised of encoders and Tr-resnet. The encoders consist of two convolutional neural networks E_c, E_a, whose targets are to extract image contents and attributes information respectively. E_c obtains 512 × 16 × 16 high-level semantic content features about source image and E_a evaluates the attribute label as the basis of label continuity operation.

The Tr-resnet concatenates the content feature from E_c and difference attribute label $l^* = l_f - E_a(x_r)$ (it is resized to the same resolution as the content feature) to construct a new feature map. Then Tr-resnet takes the new feature map as input to generate reconstructed images ($l_f = l_r$) or images with specific attributes. For the purpose of selective use of attribute information and original image information, Tr-resnet incorporates residual structure into upper convolutional layers and constructs the Tr-resnet block. The Tr-resnet block structure is shown in Fig. 2 (top right).

The discriminator D consists of a series of convolution layers, and it shares parameters with the classifier C (except for the last layer). The source image and generated image are both used as the input of discriminator and classifier. It is assumed that the image attribute label is n dimension. And the output vector of classifier is $n+1$ dimension. The first dimension is used to distinguish whether the attribute is separable or not and the remaining n dimension vector corresponds to the n-dimensional attributes of the images. By referring to the method of loss function in target detection (Redmon et al., 2016), the output vector of the generated image only takes the first dimension for loss function operation, and the other dimensions are expressed as none during training classifier stage. The detail is shown in Fig. 2 (bottom row).

3.5. Loss function

Adversarial loss Similar to other models, the work uses GAN to ensure the generated images fine result in quality. In order to stabilize training, our adversarial loss adopts WGAN-GP (Gulrajani et al., 2017).

$$L_D = -(E_{x_r \sim P_{data}} D(x_r) - E_{x_f \sim P_g} D(x_f)) + \lambda E_{x'}[(\|\nabla_{x'} D(x')\|_2 - 1)^2] \tag{7}$$

$$L_G = -E_{x_r \sim P_{data}, G(x_r, l_f) \sim P_g} D(G(x_r, l_f)) \tag{8}$$

where x' is obtained by the linear sampling between the original image and the generated image and λ is the scale of gradient penalty and sets to 10. L_D and L_G respectively represent the general adversarial loss about discriminator(D) and generator(G). Generator(G) is composed of encoder E_c, E_a (representing content encoder and attribute encoder respectively) and Tr-resnet (T). The relationship between G and E_c, T, E_a is as follows:

$$G(x_r, l_f) = T(E_c(x_r), l_f - E_a(x_r)) \tag{9}$$

Reconstitution loss StarGAN reconstructs the original images by means of cycle consistency loss, which will increase the lack of image generation during the cycle. In contrast, ClsGAN uses the attribute difference label vector $l = l_r - E_a(x_r)$ and then directly takes the label and content features into the Tr-resnet to reconstruct the image. The reconstruction loss function is as follows:

$$L_{rec} = \|x_r - T(E_c(x_r), l)\|_1 \tag{10}$$

where the L_1 norm is used to suppress blurring of reconstruction images and maintain clarity.

Object model Considering formula (5), (7), the target loss functions of joint training discriminator D and classifier C can be expressed as:

$$\min_{CD} L_{CD} = \lambda_0 L_D + \lambda_1 L_{C_d} \tag{11}$$

The objective function of the generator is comprised of adversarial loss L_G, attribute classification adversarial loss L_{C_g}, reconstruction loss L_{rec} and attribute continuity loss L_a:

$$\min_{G} L_{all} = L_G - \lambda_2 L_{C_g} + \lambda_3 L_{rec} + \lambda_4 L_a \tag{12}$$

where L_{C_d}, L_{C_g} denote attribute classification adversarial losses of classifier and generator, which are mentioned in Section 3.2. $\lambda_0, \lambda_1, \lambda_2, \lambda_3$ and λ_4 are model tradeoff parameters.

4. Experiments

Adam optimizer is adopted to train the model, and its' parameters are set to $\beta_1 = 0.5, \beta_2 = 0.999$. The learning rate of the first 10 epoch is set as 2×10^{-4}, and it is linearly attenuated to 0 at the next 10 epoch. In all experiments, the parameters are $\lambda_0 = 4, \lambda_1 = 3, \lambda_2 = 1, \lambda_3 = 20$ and $\lambda_4 = 1$. All experiments are both performed in a Pytorch environment, with training on a single NVIDIA TESLA V100. Source code can be found at https://github.com/summar6/ClsGAN.

4.1. Facial attribute transfer

Dataset This work adopts CelebA (Liu et al., 2015) dataset for training and testing of facial attribute editing. The CelebA dataset is a large face dataset, which contains more than 200,000 images of celebrities' faces and 40 facial attributes. In this paper, the last 2000 images of the dataset are set as the test set, and the remaining images are all used as the training set. The model first performs center cropping the initial size 178×218 to 170×170, then resizes them as 128×128 for training and test images.

13 of the 40 attributes are selected for attribute transfer in the paper, which are "Bald", "Bangs", "Black Hair", "Blond Hair", "Brown Hair", "Bushy Eyebrows", "Eyeglasses", "Gender", "Mouth Open", "Mustache", "No Beard", "Pale Skin" and "Age". These attributes already cover the most prominent of all attributes.

Qualitative Assessment The work compares the proposed ClsGAN with StarGAN, AttGAN, STGAN in terms of performance of facial attribute transfer. As can be seen in Fig. 3(a), while StarGAN firstly achieves the multi-attribute editing using a single model, it is still limited in manipulating large range attribute. For example, there are obvious blurs and artifacts when editing Bald and Bangs attributes, which makes the images look unrealistic. This may be because it is difficult to take full advantage of the attribute information and other content information by only convolution-residual structure. AttGAN performs better on the attribute editing and the facticity, but the results contain some differences in background compared with the original images while ClsGAN has a higher degree of restoration in the aspects of background color and skin color (see Fig. 3(b)). In addition, when performing large range editing or additive attribute using AttGAN, e.g., Bangs, Gender, Eyeglasses and Mustache, there are some blurs and artifacts (see Fig. 3). One possible reason is that the model still lacks strong implementation capability about attribute information only using the skip-connection technique of encoder–decoder and classification loss. Compared with StarGAN and AttGAN, ClsGAN accurately edits all of the attributes (global and local attributes), which credits to the applying of attribute adversarial classifier. It can be observed from Fig. 3, our results also look more normal and realistic, which benefits from Tr-resnet.

As can be seen in Fig. 4 (where Black-h, Blond-h and Brown-h represent Black-hair, Blond-hair and Brown-hair), both results of STGAN and ClsGAN can accurately edit the attribute and generate more realistic images, however, the images using STGAN are still likely to be insufficiently modified and show blurs (when editing attribute Bald) while ClsGAN has a more excellent performance. About reconstructive performance (Fig. 5), the images show more consistency with original images regardless of the content, background and other aspects, compared with competitive models.

Table 1
Quality and reconstruction performance of the comparison methods on facial attribute editing task.

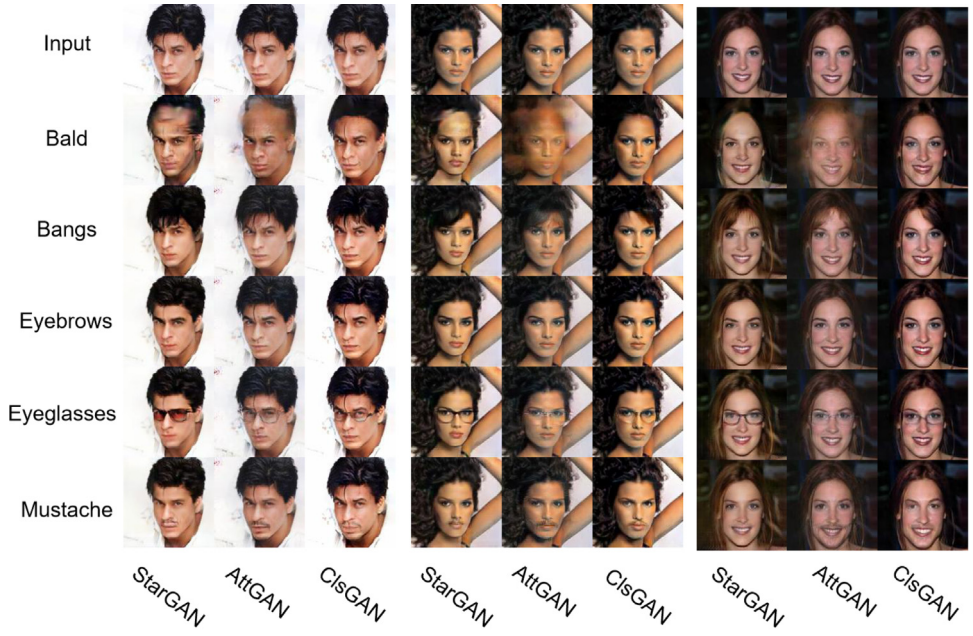
Method	StarGAN	AttGAN	STGAN	ClsGAN
FID	7.9	7.1	6.1	**6.09**
SSIM	0.56	0.8	0.92	**0.94**

As for attribute continuity, the work not only tests the synthetic images with binary attribute, i.e., with(1) or without(0), but also assumes the attribute value is continuous and may be bigger than 1. As can be seen in Fig. 6, the transfer images about different values of a single attribute are all presented to analyze the effects of attribute approximation method, which lists the editing images of ClsGAN with attribute labels of 0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6 respectively. It can be observed that the performance of the attribute gradually increases with the value that is large, which indicates that attributes have continuity. The transformation effects and image facticity are excellent from a visual point of view.

Quantitative evaluation The performance of generated images mainly needs to focus on three aspects, i.e., image quality, reconstruction accuracy and transfer accuracy. ClsGAN's purpose is to maintain the balance between quality and accuracy. The images generated by the competitive methods recently are either of low quality and high accuracy, or of low accuracy and high quality. The images from the competitive methods are generated using officially published code (StarGAN) or using a trained model (AttGAN, STGAN). Comparing to StarGAN (FID 7.9), ClsGAN greatly improves the image quality with the FID 6.09 while maintaining relatively high conversion efficiency (average 0.66). The method enhances the attribute accuracy comparing with AttGAN (average 0.637) and STGAN (average 0.59) and the image quality is better than STGAN (FID 6.1). Meanwhile, our model effectively yields transfer images to some special attributes, which is not easy to convert in other methods, such as Mustache (0.523), the result benefits from the Atta-cls.

ClsGAN utilizes two metrics, FID and SSIM, to evaluate the image quality and the similarity between reconstructed and original images, respectively. The FIDs of ClsGAN and the competitive methods are all shown in Table 1, where the test dataset of ClsGAN is adapted as the input to randomly edit 10,000 images for evaluating image quality. Our method is superior to StarGAN, AttGAN and STGAN in image quality with the FID 6.09, in which the result benefits from the application of Tr-resnet. Furthermore, the reconstruction rate outperforms other methods, which improves by 2 percentage points to 94% comparing with STGAN. It is also seen in Fig. 5 that the reconstructed images yielded by our method are more consistent with source images in each aspect (background, details, etc.) than other models.

As for the classification accuracy, the training set of CelebA dataset is adopted to train a classifier for 13 attributes and attain the average accuracy of 93.83% in the test set. Then the pre-trained classifier is used to test the transfer accuracy of different models between 2000 synthetic images. In order to compare the editing ability of each model, the conversion rate of 13 attributes is listed in the form of a bar chart. As can be seen in Fig. 7, the performance of StarGAN is outstanding in hair color transfer. A



(a) local editing

(b) global editing

Fig. 3. Face transfer images on CelebA dataset between StarGAN, AttGAN and ClsGAN. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

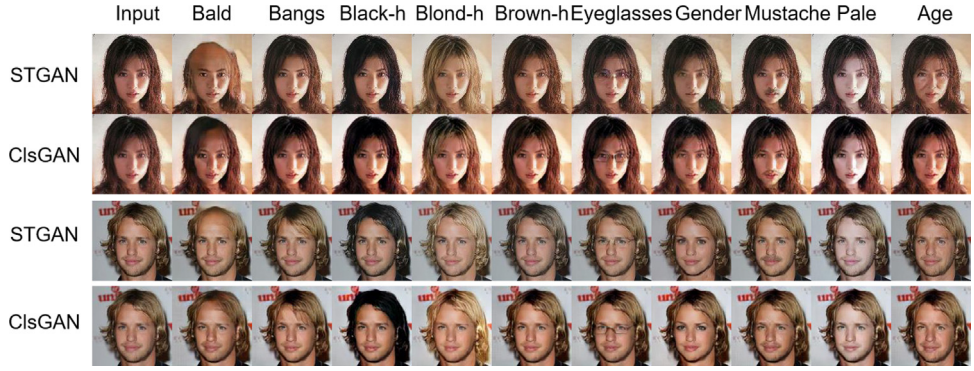


Fig. 4. Face transfer images on CelebA dataset between STGAN and ClsGAN. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 5. Face reconstructive images on CelebA dataset between different models.

possible reason is that the hair color is more prone to be affected by source attribute information owing to ease of color conversion, while the StarGAN avoids the additional introduction of source attribute information by using no skip-connection layers. Comparing with STGAN, ClsGAN attains significant gain in attribute Bangs (0.689), Black_hair (0.8325) and Mustache (0.523), which profits from the Atta-cls method. The transfer accuracy of other attributes is also relatively remarkable comparing with AttGAN and STGAN and has average transfer 0.66. Although the accuracy is slightly lower than StarGAN (average 0.74), image quality (FID 6.09) is much better than StarGAN (FID 0.79). It implicates that our attribute adversarial classifier which implements attribute generation in an adversarial way is effective in enhancing the transfer accuracy. The accuracy of attribute Pale is relatively poor in numerical terms, which is likely that the pre-trained classifier cannot recognize the image with a lighter conversion effect. However, it is acceptable to the editing effect visually (see Fig. 4).

4.2. Seasons and artistic styles transfer

Since the objective of style transfer is the same with attribute editing to some extent, ClsGAN is also implemented to realize the style transformation task. The method is employed on a season dataset and a painting dataset. Seasonal images are from the Unplash website, where the numbers of images of different seasons are: spring (29 343), summer (23 395), fall (7630) and winter (13 433). The painting images mainly come from the wikiart website, and ClsGAN achieves the mutual transformation between four styles and photographs. The number of images is Monet: 1050, Cezanne: 582, VanGogh: 1931, Ukiyo-e: 1372, Photograph: 4674. The photographs are downloaded from Flickr and use landscape labels, and are all resized as 256 × 256.

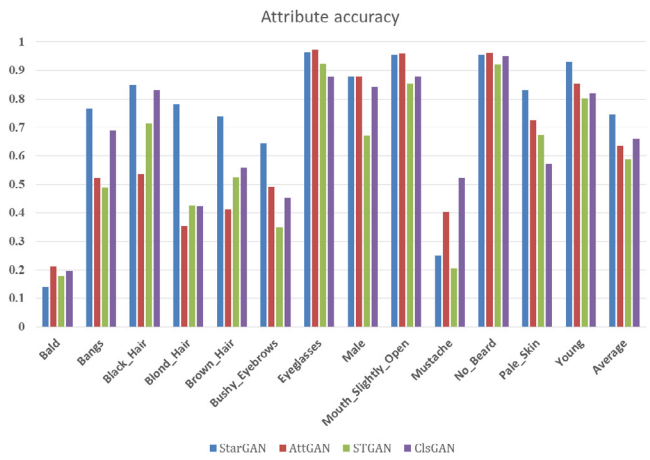


Fig. 7. The attribute accuracy about StarGAN, AttGAN, STGAN and ClsGAN. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

It can be seen from Fig. 8 that the result about the artistic and seasonal transfer is acceptable, but there are some artifacts in some synthetic images (the second image in the first row of (a), the last column of (b)). One possible reason is that it is limited when editing the large range texture and color using a single model. On the other hand, the attribute editing model may not be able to balance the effect between image quality and attribute transformation, because it needs to pay more attention to a lot of texture information. However, ClsGAN is a potential model which still deserves to be further explored and promoted.

4.3. Ablation study

In this part, the roles of Tr-resnet and attribute adversarial classifier (Atta-cls) are investigated. Concretely, four different combinations are considered: (i) ClsGAN: the original model; (ii) ClsGAN-conv: substituting Tr-resnet with the convolution network in the decoder; (iii) ClsGAN-conv-res: adopting the residual technique to learn the convolution in ClsGAN-conv; (iv) ClsGAN-oricla: adopting the original classifier which is used in StarGAN, AttGAN and STGAN instead of the adversarial classifier. Fig. 9 shows the results of different variants.

Tr-resnet vs its variants In Fig. 9 (row(1), row(2), row(3)), the results about Tr-resnet and its variants are shown (where Black-h, Blond-h and Brown-h represent Black-hair, Blond-hair and Brown-hair). It can be seen that the images outperform the other

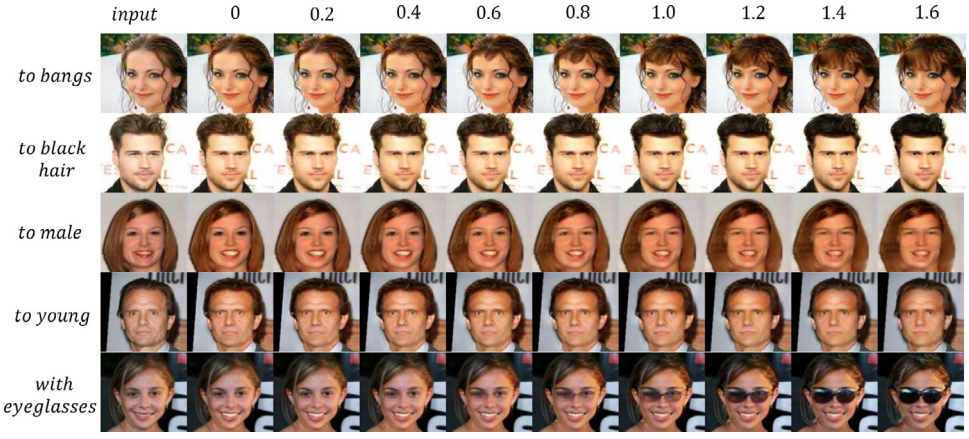


Fig. 6. Interpolation results for facial attributes on CelebA dataset by employing our model. Values among 0–1.6 are the label values about the attribute.

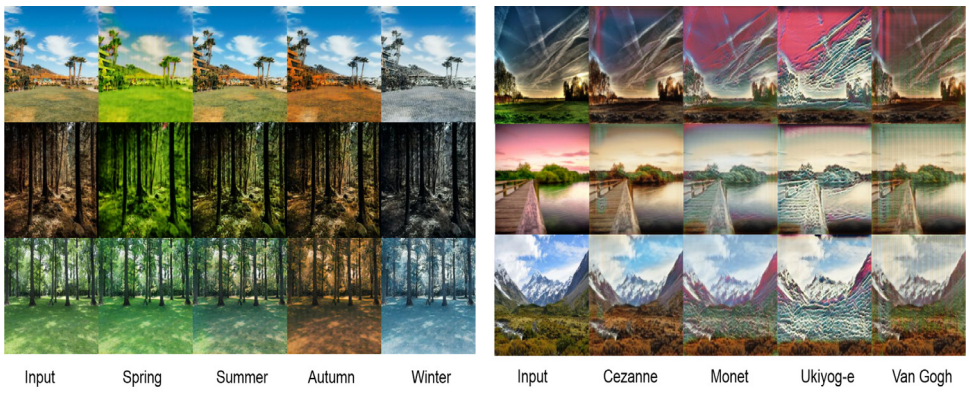


Fig. 8. The 256 × 256 transfer images about season dataset and painting dataset. Please zoom in for better observation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

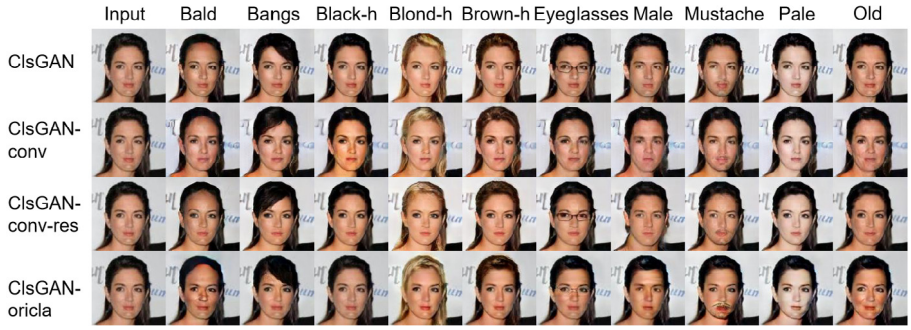


Fig. 9. Face transfer results on four different combinations. Please zoom in for better observation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

combinations using Tr-resnet. Compared with ClsGAN, the results of ClsGAN-conv are undesirable in some attributes, e.g., Eyeglasses, Old. The situation implies that it is insufficient to attain the necessary information from attributes label only using convolution operation. The images generated by ClsGAN-conv-res are relatively acceptable but the global editing contains the unnecessary changes (such as Male editing alters the hairstyle in the third row). As can be seen, the transfer images of ClsGAN are more accurate in all attributes and the quality has better performance.

Atta-cls vs original classifier Compared with Atta-cls, the results (see row(4)) of using the original classifier look more unrealistic accompanied with artifacts and blurriness. One available reason is that only using the original classifier is likely to affect the gain of the necessary information from original images and attributes label, so as lower performance of photo-realistic and attribute accuracy. In this paper, the adversarial technique is employed to motivate the classifier to learn detailed attribute information, thus it improves the quality of images by optimizing the generator. From row(1) of Fig. 9, it can be seen that the results of ClsGAN look more photo-realistic and natural compared with the original classifier method.

5. Conclusion

In this paper, the work first analyzes the constraint problem between image attribute transfer and quality about attribute editing and proposes the ClsGAN model by incorporating the upper convolution residual network (Tr-resnet) and attribute adversarial classifier (Atta-cls). About Tr-resnet, the upper convolution structure is applied with residual technique to select the desired information and avoid interference with resource attribute information. What is more, Atta-cls is presented to enhance the attribute transfer accuracy of the image, which is inspired by the spirit of generated adversarial network. The attribute adversarial classifier can selectively find necessary information about attributes and then optimize the generator in an adversarial way. At the same time, an approximation between the source label and attribute feature vector (which is generated by style encoder) is made to meet the requirement of label continuity. Experiments and ablation studies both demonstrate the great effectiveness of ClsGAN in attribute 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 was primarily supported by National Key R&D Program of China (No. 2018YFC1604000) and Foundamental Research Funds for the Central Universities (Program No. 2662017JC049) and State Scholarship Fund (No. 261606765054).

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214–223).

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. In *International conference on learning representations*.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8789–8797).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in neural information processing systems* (pp. 5767–5777).

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2017). Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing, 28*, 5464–5478.

Hinton, G. E., & Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In *Advances in neural information processing systems* (pp. 3–10).

Hinton, G. E., & Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In *Advances in neural information processing systems* (pp. 3–10).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. In *International conference on learning representations*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4401–4410).

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4401–4410).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. In *International conference on learning representations*.

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., & Matas, J. (2018). Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8183–8192).

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning* (pp. 1558–1566).

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., & Wang, Z. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4681–4690).

Li, X., Hu, J., Zhang, S., Hong, X., Ye, Q., Wu, C., & Ji, R. (2019). Attribute guided unpaired image-to-image translation with semi-supervised learning. arXiv preprint arXiv:1904.12428.

Li, M., Zuo, W., & Zhang, D. (2016). Deep identity-aware transfer of facial attributes. arXiv preprint arXiv:1610.05586.

Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., & Wen, S. (2019). STGAN: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3673–3682).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730–3738).

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *International conference on learning representations*.

Nowozin, S., Cseke, B., & Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems* (pp. 271–279).

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International conference on learning representations*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Schmidhuber, J. (2020). Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991). *Neural Networks, 127*, 58–66.

Wu, P.-W., Lin, Y.-J., Chang, C.-H., Chang, E. Y., & Liao, S.-W. (2019). RelGAN: Multi-domain image-to-image translation via relative attributes. In *2019 IEEE/CVF international conference on computer vision* (pp. 5913–5921).

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492–1500).

Xie, D., Yang, M., Deng, C., Liu, W., & Tao, D. (2019). Fully-Featured Attribute Transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*(pp. 3431–3440).

Zhou, S., Xiao, T., Yang, Y., Feng, D., He, Q., & He, W. (2017). Genegan: Learning object transfiguration and attribute subspace from unpaired data. *In British Machine Vision Conference*.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).