EURASIP Journal on Image and Video Processing

RESEARCH Open Access

# Facial attribute-controlled sketch-to-image translation with generative adversarial networks

Mingming Hu[1*] and Jingtao Guo[2]

**Abstract**

Due to the rapid development of the generative adversarial networks (GANs) and convolution neural networks (CNN), increasing attention is being paid to face synthesis. In this paper, we address the new and challenging task of facial sketch-to-image synthesis with multiple controllable attributes. To achieve this goal, first, we propose a new attribute classification loss to ensure that the synthesized face image with the facial attributes, which the users desire to have. Second, we employ the reconstruction loss to synthesize the facial texture and structure information. Third, the adversarial loss is used to encourage visual authenticity. By incorporating above losses into a unified framework, our proposed method not only can achieve high-quality sketch-to-image translation, but also allow the users control the facial attributes of synthesized image. Extensive experiments show that user-provided facial attribute information effectively controls the process of facial sketch-to-image translation.

**Keywords:** Facial attribute editing, Face sketch to image translation, Generative adversarial networks (GANs), Facial attribute classifier

## 1 Introduction

Due to the wide application of the face sketch image to color image translation in public security system and digital image processing industry, it has become an important research topic in the field of computer vision and deep learning [1–5]. In the field of public security systems, most local surveillance systems are not perfect, so police officers are often unable to obtain high-quality, complete color images of suspects which often leads to police unable to confirm the identity of the suspect by comparing them with database images. In response to this situation, the police can only compare the image of the database with the witness's sketch image of the suspect, but the image comparison success rate is extremely low due to the huge difference between the sketch image and the color image. It has brought great difficulties to the arrest of criminal suspects. In this case, the study of the translation of the face sketch image to a color image has greatly helped the police to confirm the identity of the suspect. In the field of digital entertainment, many people want to have software, which can convert their own sketches into color images while transforming their hair color or skin color into their favorite color. In those cases, a method is necessary, which not only can generate high-quality reconstructed image from sketch to image translation, but also can control the facial attribute changing during the sketch to image translation.

Deep learning and related network models have become the focus of research, especially the rapid development of the generative adversarial networks (GANs) [6–8] has greatly promoted the research of image to image translation, it improved both the quality and the efficiency of image to image translation work [9–18]. Different from the traditional dictionary-based traditional network model, Zhang et al. [19] proposed an end to end deep convolutional neural network (CNN) architecture for an image to image translation research. Isola et al. [20] attach an additional condition $y$ to the traditional GANs and use it as part of the input layer to control the mapping between the input image and the generated image during image translation.

In recent years, sketch to image or image to sketch translation have been widely used in digital entertainment

* Correspondence: 457835932@qq.com
[1] Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
Full list of author information is available at the end of the article

field and law enforcement [21–29]. Zhang et al. [21] proposed a face sketch generator based on the compositional model, which can address the problem of generating sketches with over-smoothing effects in existing face sketch synthesis methods. The proposed method can obviously reduce the high frequency loss and has a good performance in sketch image synthesis. To have a better performance in sketched face identification, Zhang et al. [22] proposed a novel dual-transfer face sketch-image synthesis architecture, which is composed of an inter-domain transfer process and an intradomain transfer process. The inter-domain transfer is used to ensure the recovery of common facial structures during face sketch synthesis; the intra-domain transfer is used to suppress the loss of identity-specific information. To synthesis face sketch with fine face detail features, Zhang et al. [23] proposed a bionic face sketch generator, which is consisted of the coarse part, the fine part, and the finer part. Through the combination of three parts, the proposed method can synthesize a face sketch with delicate and face detail features. Zhang et al. [24] proposed a novel face sketch synthesis architecture, which combining the probabilistic graphical model (NPGM). The mapping of the pixels of generated sketch and the candidates selected from the training data is considered in this method, which is the key of the generated face sketch with fine face detail features. Current existing face sketch synthesis methods are vulnerable to lighting variations and cannot generate satisfactory face sketches with lighting varies, to address this problem, Zhang et al. [25] proposed a novel cascaded face sketch synthesis architecture, which is consisted of a multiple feature generator and a cascaded low-rank representation. The multiple feature generator can extract a photo detail features with various illuminations, and the cascaded low-rank representation can reduce the gap between the generated face sketch and the corresponding sketch; the experiment results show that the method have a better performance than the existing methods. Wang et al. [26] proposed a simple but novel face sketch synthesis method instead of on-line K-NN search method, which can obviously improve the face sketch synthesis efficiency. The experiment results show that the proposed method has a better performance in synthesis quality and synthesis efficiency. Wang et al. [27] proposed a face sketch synthesis method based on Bayesian architecture, which considered the constraint both in the neighbor selection model and in the weight computation model. Compared with the existing method, the proposed method based on Bayesian architecture has better performance both in subjective perceptions and objective evaluations. In existing face sketch synthesis methods, it is difficult to select the accurate neighbor during the face sketch synthesis, the K-nearest neighbor (KNN) matching algorithm is always used in existing methods. Wang et al. [28] proposed a simple but effective neighbor selection algorithm, which is named anchored neighborhood index (ANI). Experiment results show that the proposed method have a better performance in face sketch quality and face recognition accuracy. Most existing face sketch synthesis methods use a unidirectional feedforward mapping during the synthesis process, G: X→Y or F: Y→X, the utilization of both two opposite mappings is lacking in existing synthesis face sketch methods. Zhu et al. [29] proposed a collaborative architecture for face sketch synthesis, which can use both two opposite mappings, the proposed method can constrain the two opposite mappings, thus obtaining a better performance in generating face sketch image.

However, current proposed methods, neither the CNN-based nor the GAN-based methods [19, 30–34] are able to control the change of facial attributes during the face sketch to image translation. They only pay attention to the fit between the input image and the output image, so that the converted output image looks more natural and realistic. However, in practical application research, relevant researchers may wish to control the facial attribute to become what they desired during the image translation process. Moreover, they did not provide some loss learning functions to preserve the detailed features of the input image. As a result, the generated image is severely distorted or unnatural.

To solve those problems which are existed in existing methods, a convolutional generative network with encoder-decoder architecture is proposed to achieve the facial attribute-controlled face sketch to image translation. Our proposed network architecture not only generates high-quality output images but also can control the changes of facial attribute without affecting other facial details. Our network architecture consists of three parts: generator, discriminator, and facial attributes classifier. Based on the encoder-decoder architecture, the decoder implements facial attributes controlled during face sketch to image translation by decoding potential features extracted by the encoder conditioned on the desired facial attributes. The training process of the network architecture is as follows: First, we guide the generator to generate an image with the desired facial attributes continuously by categorizing the facial attribute classification loss; Second, we use reconstruction loss learning to preserve facial attributes and other face feature details during face image reconstruction; At last, we use adversarial loss learning to enhance the visual authenticity of the generated image. These three loss learning functions are perfectly integrated into our proposed network architecture, the proposed network architecture not only can generate an output image corresponding to the input facial image detail feature, but also allows us to control the changes of the facial attribute during the face sketch to image translation.

Figure 1 shows an example of the face sketch to image translation with controllable facial attributes through our proposed network architecture.

Our network architecture is similar to the AttGAN [35], which makes a great improvement in facial attribute editing, but we found two shortages during our experiments with the network architecture of the AttGAN: First, the detail features of synthetic image are severely distorted and blurry; second, we cannot control the facial attribute changing obviously during the sketch to face image translation. With the limit of the AttGAN, we make several important modifications with the network architecture of the AttGAN: First, to better control the changes of facial attribute during the sketch to image translation, we add the desired facial attribute b at the beginning of down-sampling layer (the encoder); the method can extract the characteristics of facial attribute better; Second, we introduce some residual blocks to reserve the detail features of the generated image.

The contributions of our method are as follows:

- First and foremost, we make some important modifications on the network of the AttGAN; our proposed network architecture has a better performance in controlling the changes of facial attribute during the sketch to image translation.
- Our method can generate higher quality reconstructed images during the sketch to image translation.

The rest of the paper is organized as follows: In Section 2, we will introduce the related work about the image to image translation and facial attribute editing research. Section 3 described the proposed novel network architecture which combines GANs and encoder-decoder architectures to complete the face sketch to image translation with controllable facial attributes. In Section 4, we will show the experimental results of our proposed method on the CelebA dataset. We will summarize the main work in this paper at last.

## 2 Related work

### 2.1 Generative adversarial networks

GANs [8, 36] are composed of two networks, a generator network and a discriminator network. The generative adversarial networks use the confrontation of two neural networks as the training criterion and use back propagation to train. In the training process, the traditional Markov chain method is abandoned, and there is no complicated variational lower bound. Therefore, the training efficiency and training difficulty of generating model have been improved. Besides, the GANs can directly sample and infer new samples in the encoding process, which improves the coding efficiency of the original samples significantly. The GANs have been widely used in image generation [8, 19, 25, 32, 37, 38], image translation [5, 33], and image synthesis [11, 12]. Moreover, these researches have got remarkable achievements.

The distribution of a real image $x$ is denoted by $p_{data}(x)$, the distribution of input data denoted by $G(z)$, and $D(x)$ denoted as mapping the input noise $z$ into the input data. The optimal-objective function for GANs is defined in the form of minmax as Eq. (1):

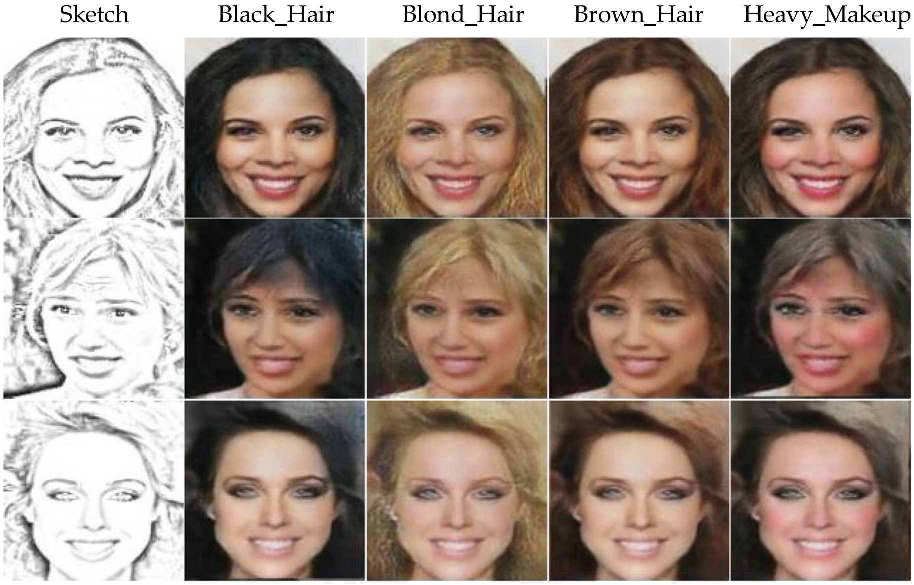

**Fig. 1** Facial attribute-controlled sketch to image translation results with our proposed method on CelebA dataset. The first column shows input face sketch, and others are generated with a controllable facial attribute by our proposed method

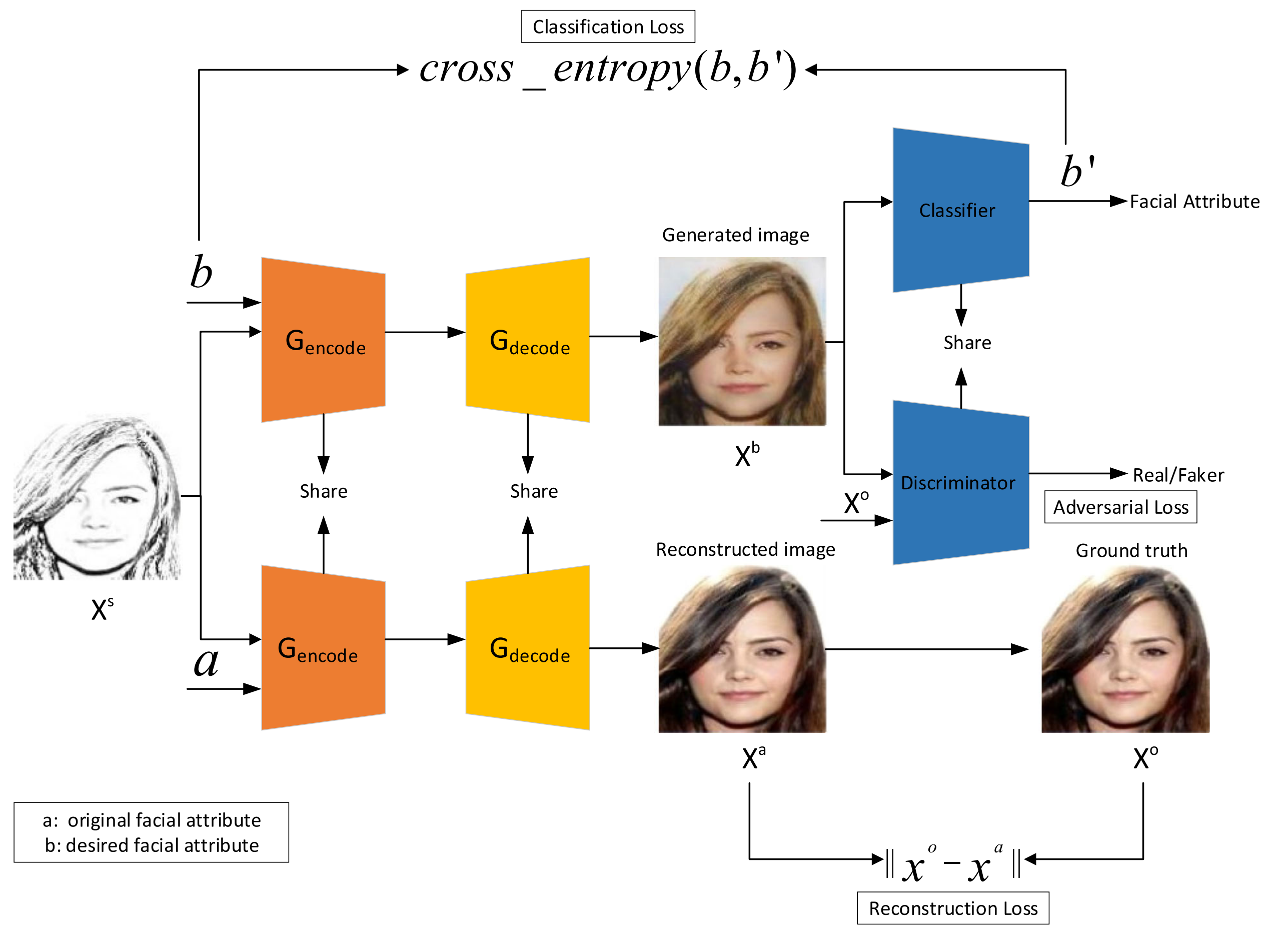

**Fig. 2** Overview of our proposed method, which consists of three main models at training: the generator, the generator consists of an encoder and a decoder, the discriminator, and the attribute classifier. The facial attribute classifier ensures the desired facial attribute manipulation on the input image. The generator is used to preserve the input image detail features. The discriminator is employed for a visually realistic generation

**Table 1** Network architecture of our proposed method

| Generator | | Discriminator | Attribute classifier |
|---|---|---|---|
| Conv(64, 5, 1) | | Conv(64, 3, 2) | |
| Conv(128, 3, 2) | | Conv(128, 3, 2) | |
| Conv(256, 3, 2) | | Conv(256, 3, 2) | |
| Residual blocks | Conv(256, 3, 1) | Conv(512, 3, 2) | |
| | Conv(256, 3, 1) | Conv(1024, 3, 2) | |
| | Conv(256, 3, 1) | FC(1024) | FC(1024) |
| | Conv(256, 3, 1) | FC(1) | FC(7) |
| | Conv(256, 3, 1) | | |
| | Conv(256, 3, 1) | | |
| DeConv(128, 3, 2) | | | |
| DeConv(64, 3, 2) | | | |
| DeConv(3, 5, 1) | | | |

$$\min_{G}\ \max_{D} E_{x\sim p_{\text{data}}(x)}[\log(D(x))] + E_{z\sim p_{z}(z)}[\log(1-D(G(z)))] \tag{1}$$

In recent years, many kinds of network models have been derived from the original GANs. Conditional generative adversarial networks (cGANs) [39–41] realize the traditional GANs by transferring additional information $y$ to discriminant model and generation model. Deep convolutional generative adversarial networks (DCGAN) [36, 42] improves the stability of network training and the quality of generating results by changing the network structure of traditional GANs. Wasserstein GAN [37, 43] optimizes the traditional GANs from the perspective of a loss function. Based on these network models, we propose a framework which combines GANs and encoder-decoder architecture to implement the research of face sketch to image translation with controllable facial attributes.

### 2.2 Image to image translation

The goal of an image to image translation is to learn the mapping relationship between the input image and the output image through the network model, to generate a new image associated with the input image. In recent years, many image to image translation methods based on deep learning model have been proposed by relevant researchers and many significant achievements have been achieved. Isola et al. [20] propose a cGAN method to solve existing problems in the image to image translation. The proposed network not only learns the mapping relationship between the input image and output image but also feeds back the mapping relationship through loss function. Zhang and Lin [19] propose a novel network architecture based on CNN for end-to-end face image to sketch translation. They use a new optimization objective function to maintain the features of the input image in the process of image translation. To solve the problem of unavailability of paired training data in some cases, Zhu.et al. [38] propose a network architecture (e.g., converting zebra images into horse images) which can translate input images into target images without paired training samples, and this network architecture is called Cycle GANs. However, these proposed methods are unable to control the changes of the facial attributes during image to image translation.

### 2.3 Facial attribute editing

The goal of facial attribute editing is to be able to change some of the facial attributes of the input facial image (e.g., change the hair color or skin color) without affecting other details of the facial image. Research on facial attribute editing based on deep CNN or GANs has achieved a lot of achievements. To solve the problem that the new image generated by the existing facial attribute generation models cannot maintain the input image facial details, Li et al. [44] propose an optimized network model based on deep convolutional neural network, called VGG-Face. The network model can generate a new image with the desired facial attributes by guiding the expected facial attributes or adjusting the facial attribute features of the input image. Patsorn et al. [18] introduce a deep adversarial synthesis architecture, which can generate realistic images from sketch with sparse color; besides, the method also can generate realistic images with controlled color (e.g., car color, face color). Therefore, AttGAN [35] introduces the face attribute classifier, which is used to guarantee the correct generation of the desired face attributes during the facial attribute classification. To map the input image into potential space and represent it, IcGAN [45] reverses the mapping in the cGANs by retraining the image encoder. This method allows us to re-edit or modify the facial attributes of the input image. However, in the context of facial attribute editing research, these methods lack a model which can implement the face sketch to image translation. We propose a conditional facial image translation model which can implement the face sketch to image translation with controllable facial attribute editing.

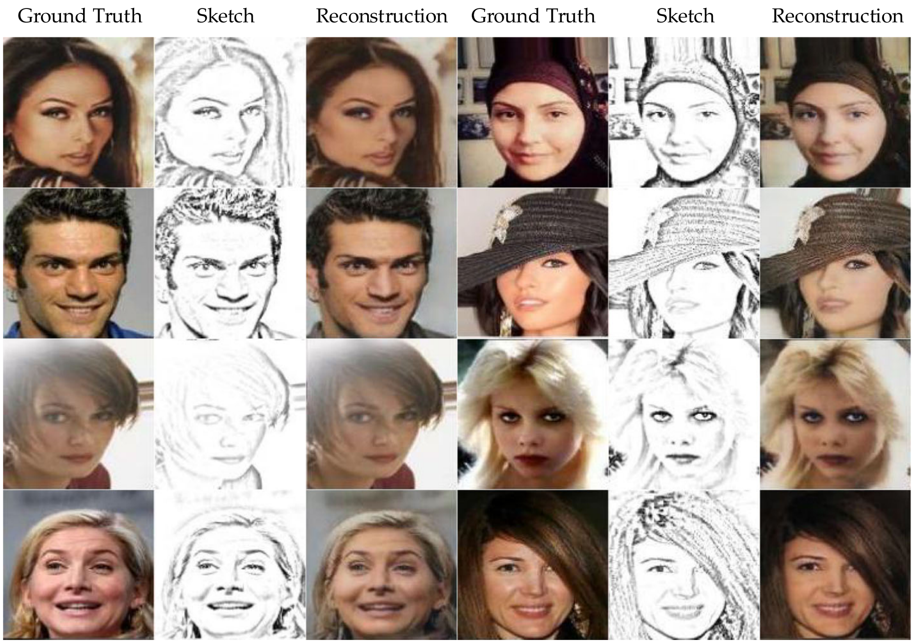

**Fig. 3** Sketch to image translation without controllable facial attribute with our proposed method on CelebA dataset. The first and fourth columns are ground truth, the second and fifth columns are input sketch, and the third and sixth columns are reconstructed image

## 3 Method

In this section, we first describe our proposed model, a network architecture to address face sketch to image translation with controllable facial attribute editing. Then, we will discuss the network architecture of proposed method. At last, details of the main loss formulations and testing formulation are described below.

### 3.1 Network architecture

Our goal is to allow the user to control changes in facial attributes during the face sketch to color image translation. As shown in Fig. 2, our network is comprised of the generator G, the facial attribute classifier C, and the discriminator D, the network architecture used in our method is shown in Table 1. The generator G consists of an encoder $G_{\text{encode}}$ and a decoder $D_{\text{decode}}$, which has two down-sampling convolutional layers, six residual blocks and two up-sampling convolutional layers, the encoder $G_{\text{encode}}$ takes the face sketch image $x^s$ and the facial attribute vector $a$ as input and produces a potential feature representation of the input image, the decoder $G_{\text{decode}}$ receives the feature representation and the desired facial attribute vector $b$ as input and outputs an image $x^b$ with desired facial attribute. The facial attribute classifier C consists of five down-sampling convolutional layers and one fully connected layer, which used to minimize the desired facial attribute classification error. The discriminator D includes five down-sampling convolutional layers and a fully connected layer, and a sigmoid function is used to predict whether the input image is real or fake.

Starting with the network architecture design of the AttGAN, we make some important modifications to

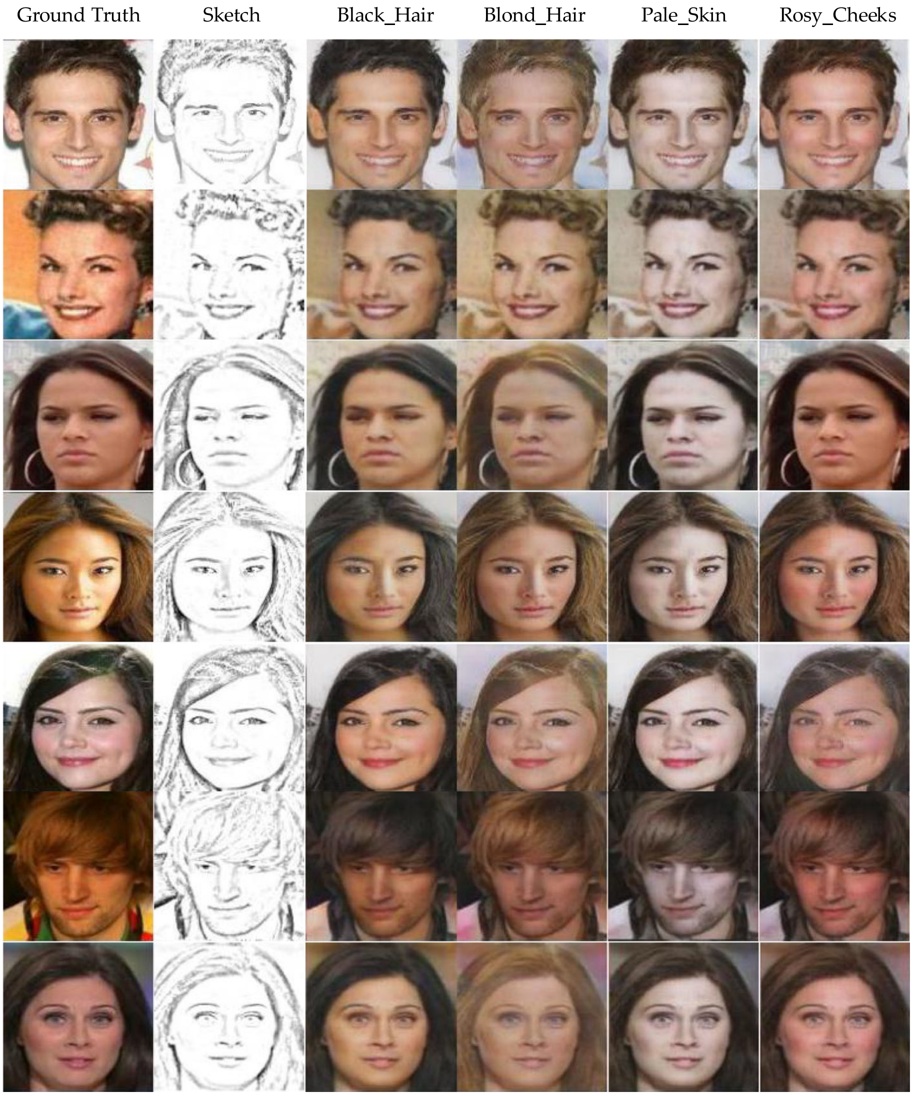

**Fig. 4** The face sketch to image transfer results in a single-controlled facial attribute with our proposed method on the CelebA dataset. The first column shows ground truth; the second column shows input images while the remaining columns show the single facial attribute transfer results

improve the quality of reconstructed image and generated image with controlled facial attribute: (1) to achieve better reconstructed results, the facial attribute vector $a$ or $b$ is inputted at the beginning of the down-sampling convolutional layer (the encoder) which is different from the AttGAN; (2) we add six residual blocks between the encoder and the decoder, which can better reserve the detail features of generated image.

### 3.2 Loss function

Our proposed network architecture mainly includes three loss learning functions: image reconstruction loss, adversarial loss, and facial attribute classification loss. We will discuss them in detail below.

The $x^a$ is denoted as the color facial image with facial attributes $a = [a1, a2, ...., an]$, its corresponding face sketch image is denoted as $\tilde{x}$, giving the face sketch image $\tilde{x}$ and desired facial attributes $b = [b1, b2, ...., bn]$. The generator $G$ can synthesize a color image $b$ with the facial attribute $x^b$.

#### *3.2.1 Image reconstruction loss*

The purpose of reconstruction loss learning is to preserve facial attributes and other feature details during the facial image reconstruction. Therefore, the decoder needs a reconstruction loss function to reconstruct the input image $\tilde{x}$ by decoding the latent representation conditioned on the original facial attributes of the input image $a$. The reconstruction loss formulated as Eq. (2):

$$L_{\text{rec}} = \left\| x^a - x^{\hat{a}} \right\|_1 \tag{2}$$

where $x^{\hat{a}} = G(\tilde{x}, a)$.

#### *3.2.2 Adversarial loss*

To make the translated image $x^b$ visually realistic and natural, the adversarial loss for the discriminator $D$ and the generator $G$ are formulated as Eqs. (3) and (4):

$$L_{\text{adv}_d} = -E\left[ \log(D(x^a)) + \log\left(1 - D\left(x^{\hat{b}}\right)\right)\right] \tag{3}$$

$$L_{\text{adv}_g} = E\left[ \log\left(D\left(x^{\hat{b}}\right)\right)\right] \tag{4}$$

where $x^{\hat{b}} = G(\tilde{x}, b)$, the generator network $G$ produces the image $x^{\hat{b}}$ conditioned on both the face sketch $\tilde{x}$ and the target facial attributes $b$, while the discriminator network $D$ tries to distinguish between a real and a generated image.

#### *3.2.3 Facial attribute classification loss*

To train the facial attribute classifier $C$ to classify color facial image attribute tags correctly, we define the summation of the binary cross entropy losses between an attribute $a_i$ and its facial attribute classification $C_i$.

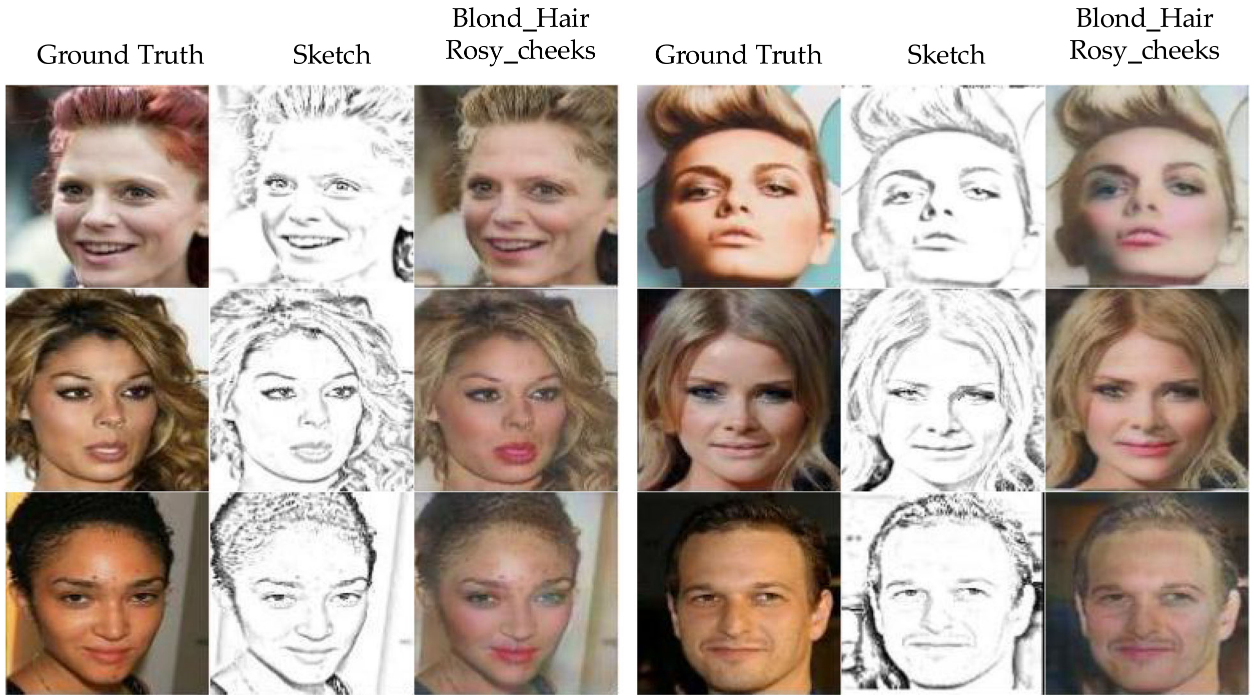

**Fig. 5** The face sketch to image transfer results with multiple controllable facial attributes ("Blond Hair" and "Rosy_Cheeks") with the proposed method on the CelebA dataset. The first and fourth columns are ground truth, the second and fifth columns are input sketch, and others are generated image

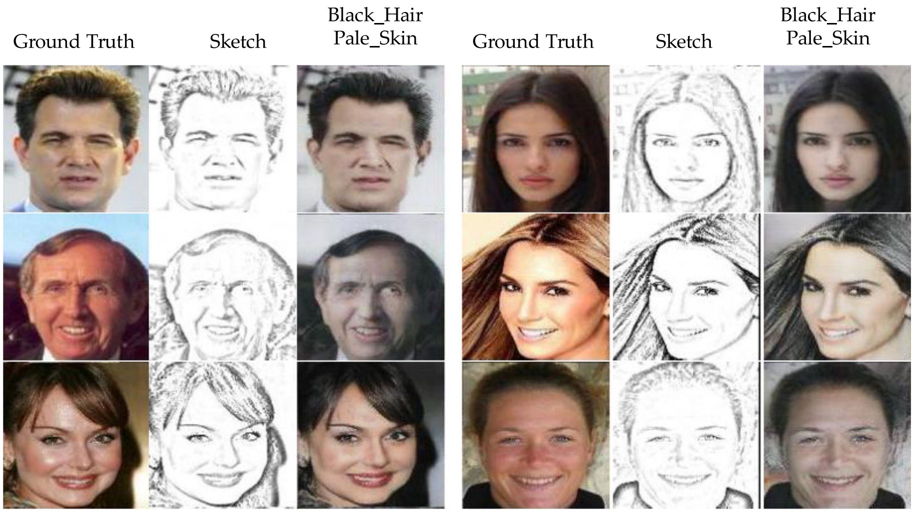

**Fig. 6** The face sketch to image transfer results with multiple controllable facial attributes ("Black_ Hair" and "Pale_Skin") with the proposed method on the CelebA dataset. The first and fourth columns are ground truth, the second and fifth columns are input sketch, and others are generated image

$$L_{\mathrm{cls}_c} = \sum\nolimits_{i=1}^{n} -a_i \log C_i\left(x^{\hat{a}}\right) - (1-a_i) \log\left(1 - C_i\left(x^{\hat{a}}\right)\right) \tag{5}$$

where $x^{\hat{a}} = G(\tilde{x}, a)$, the facial attribute classifier $C$ is trained on the face sketch $\tilde{x}$ with labeled facial attributes $a$, and $C_i(x^{\hat{a}})$ indicates the prediction of the $i$th attribute. By minimizing this objective, $C$ learns to classify a raw image $x^a$ into its corresponding original facial attributes $a$.

To optimize the generator to synthesize color images with target facial attribute, we define the summation of the binary cross entropy classification losses for training $G$:

$$L_{\mathrm{cls}_g} = \sum\nolimits_{i=1}^{n} -b_i \log C_i\left(x^{b}\right) - (1-b_i) \log\left(1 - C_i\left(x^{b}\right)\right) \tag{6}$$

here $x^{\hat{b}} = G(\tilde{x}, b)$, $C_i(x^b)$ indicates the prediction of the $i$th facial attribute for a color face image $x^b$. By minimizing this objective, the generator $G$ learns to generate $x^b$ with the desired facial attributes $b$.

To make things clear, the final objective functions to optimize the $G$, $D$, and $C$ networks are listed below:

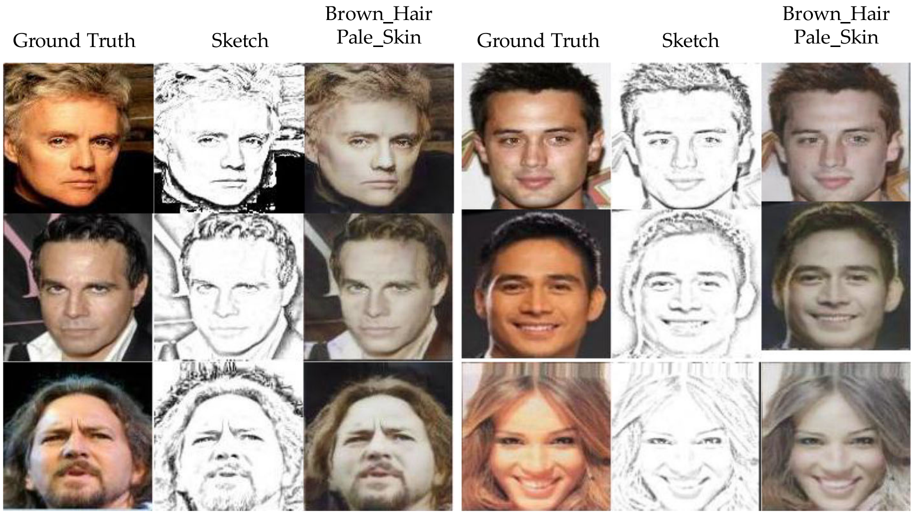

**Fig. 7** The face sketch to image transfer results with multiple controllable facial attributes ("Brown_Hair" and "Pale_Skin") with the proposed method on the CelebA dataset. The first and fourth columns are ground truth, the second and fifth columns are input sketch, and others are generated image

$$L_G = L_{\text{rec}} + \lambda_{\text{adv}} L_{\text{adv}_g} + \lambda_g L_{\text{cls}_g} \tag{7}$$

$$L_D = L_{\text{adv}_d} \tag{8}$$

$$L_C = \mathrm{L}_{\text{cls}_c} \tag{9}$$

here, $\lambda_{\text{adv}}$ and $\lambda_g$ are weights to define the importance of different losses for the generator network.

### 3.3 Testing function

To train our network model, we need to define test functions to test its performance. The generator encoder $G_{\text{encode}}$ is applied to encode face sketch $\tilde{x}$ into the latent representation $z$, denoted as Eq. (10):

$$z = G_{\text{encode}}(\tilde{x}) \tag{10}$$

then the generator decoder $G_{\text{decode}}$ is implemented to achieve the process of changing the facial attributes of sketch $\tilde{x}$ to other facial attributes $b$ by decoding $z$ conditioned on $b$, denoted as eq.(11):

$$x^b = G_{\text{decode}}(z, b) \tag{11}$$

Thus, the whole editing process is indicated as Eq. (12)

$$x^b = G_{\text{decode}}(G_{\text{encode}}(\tilde{x}), b) \tag{12}$$

## 4 Experiments

In this section, we demonstrate the ability of our proposed method to provide some face sketch to image translation results with some desired facial attributes.

### 4.1 Dataset

We train our proposed network model and evaluate its performance on CelebA [46] dataset. The CelebA dataset contains 202,599 face images of celebrities, which has more than 40 binary facial attributes. During the training process, we randomly select 20,000 images as the test set and use the rest images as training data. We convert the color image of the dataset into face sketch image as an input dataset by applying xDOG [47] filter. In our experiments, we chose seven facial attributes with strong visual impact, including “Black_Hair”, “Blond_Hair”, “Brown_Hair”, “Heavy_Makeup”, “Gray_Hair”, “Pale_Skin”, “Rosy_Cheeks”.

### 4.2 Implement details

We train our proposed network model with TensorFlow deep learning framework and execute it on Lenovo

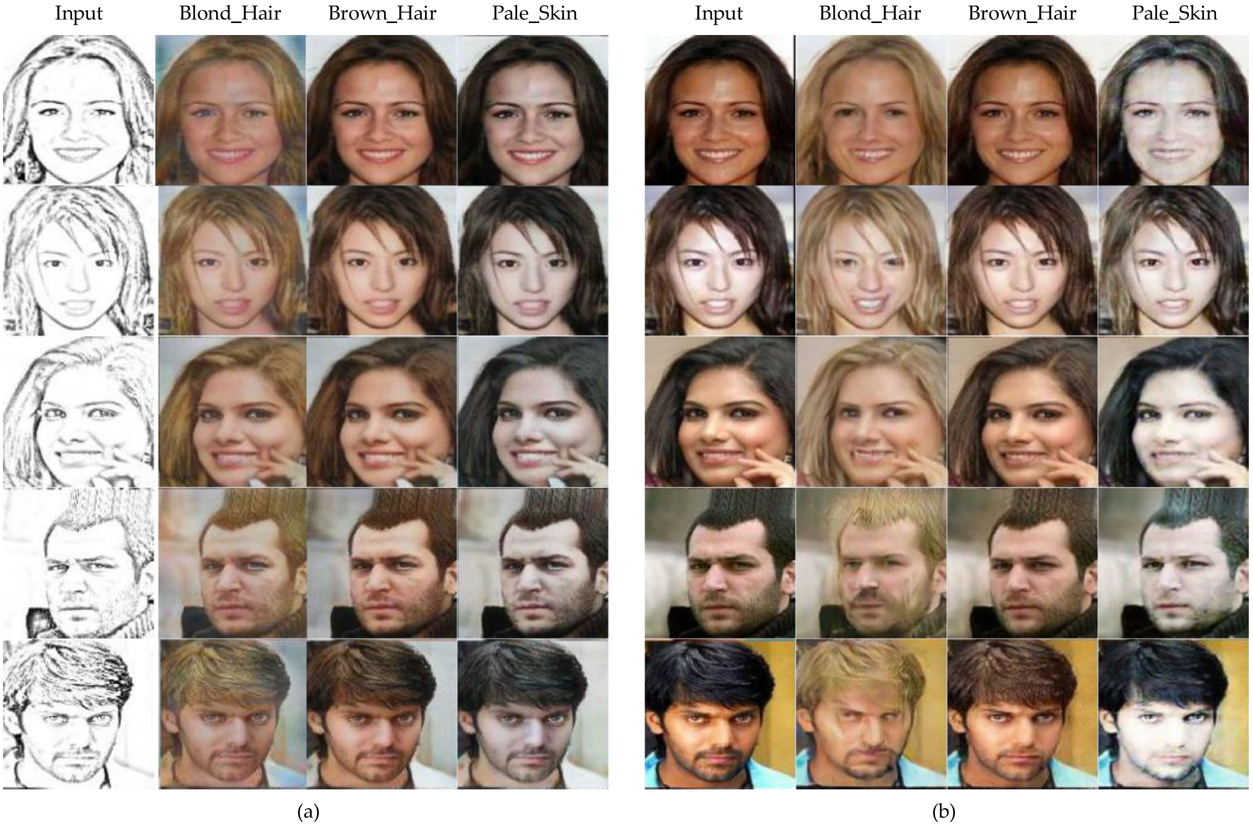

**Fig. 8 a** The image which is generated with our proposed network architecture and **b** the image which is generated with the AttGAN

ThinkCentre computer with NVIDIA 1080Ti GPU (8GB). The network model is trained with 128 × 128 face sketch, and the batch size is set to 16 during our experiment. The loss factor for Eq. (7) is set as: $\lambda_{\text{adv}} = 1$ and $\lambda_g = 15$, which is used to balance the effects of various losses.

## 5 Results and discussions

Existing image to image translation network architecture can only output an image associated with the input image detail features. In these methods, the particular image to image translation result is automatically generated by their proposed models, which do not allow us to change facial attribute during image to image translation. In contrast, our approach not only learns the mapping from an input image to output image during sketch to image translation but also allows us to change the facial attributes of the translated image.

### 5.1 Facial image translation without controllable facial attribute

To evaluate the performance of our proposed model in face sketch to image translation, we first make the experiment of face sketch to image translation without the controllable facial attribute. As shown in Fig. 3, the reconstructed face images with our proposed network architecture are more realistic and natural.

### 5.2 Sketch to image translation with a single controllable facial attribute

We make the experiment of the face sketch to image translation with manipulating a single facial attribute (“Black_Hair”, “Blond_Hair”, “Pale_Skin”, “Rosy_Cheeks”), which is related to one local key facial feature (hair or skin). As shown in Fig. 4, compared with the ground truth, the change in facial attribute of the generated face images are very noticeable; moreover, they seem to have no sense of disobedience.

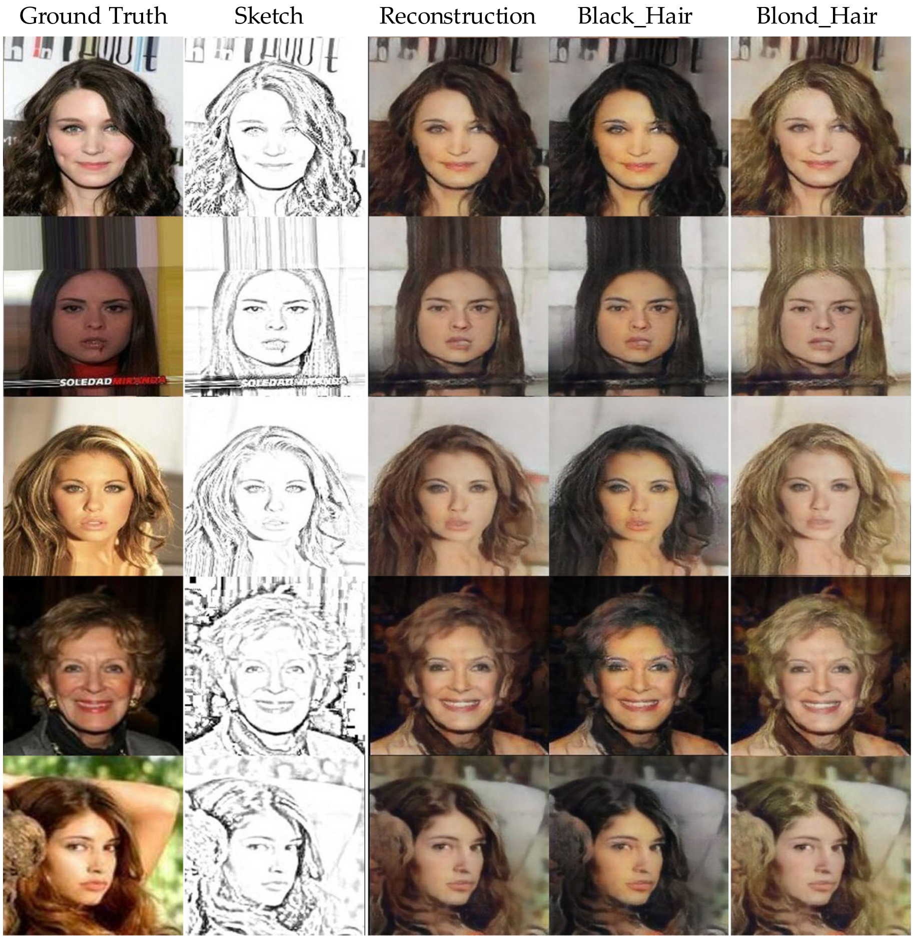

**Fig. 9** The face sketch to image transfer results in a single-controlled facial attribute with 256 × 256 image. The first column shows ground truth; the second column shows input images while the remaining columns show the single facial attribute transfer results

### 5.3 Sketch to image translation with multiple controllable facial attributes

Our proposed network architecture also can generate realistic and natural face images with multiple facial controllable facial attributes. As shown in Figs. 5, 6, and 7, the proposed network architecture has a good perform in complex combinations of facial attributes.

### 5.4 Qualitative comparisons

To evaluate the performance of our proposed network architecture, we compare our experiments with the existing method. Due to current methods cannot control the change during the sketch to image translation, we compare the results with the AttGAN [35] which aims to change the facial attribute during the image reconstruction. For fair comparison, we retrain the network model of the AttGAN in the same epoche on the CelebA data. As shown in Fig. 8, the textures of generated images are blurry with the network model of the AttGAN; in contrast, the generated images with our network architecture are more realistic and natural. Therefore, our network architecture has a more powerful modeling range from low-level textures to high-level semantic structures than the AttGAN.

### 5.5 Performance evaluation of our proposed method with high pixel image

All about experiments are conducted with 128 × 128 images. To evaluate the performance of our proposed method, we also trained our proposed network model with 256 × 256 images, the result demonstrated that the proposed method also has a good performance in high pixel image, as shown in Fig. 9.

### 5.6 Sketch to image translation on the UT Zappos50K dataset

We also make the experiment of the sketch image to color image translation on the UT Zappos50K Dataset. In this experiment, the input sketch is generated without

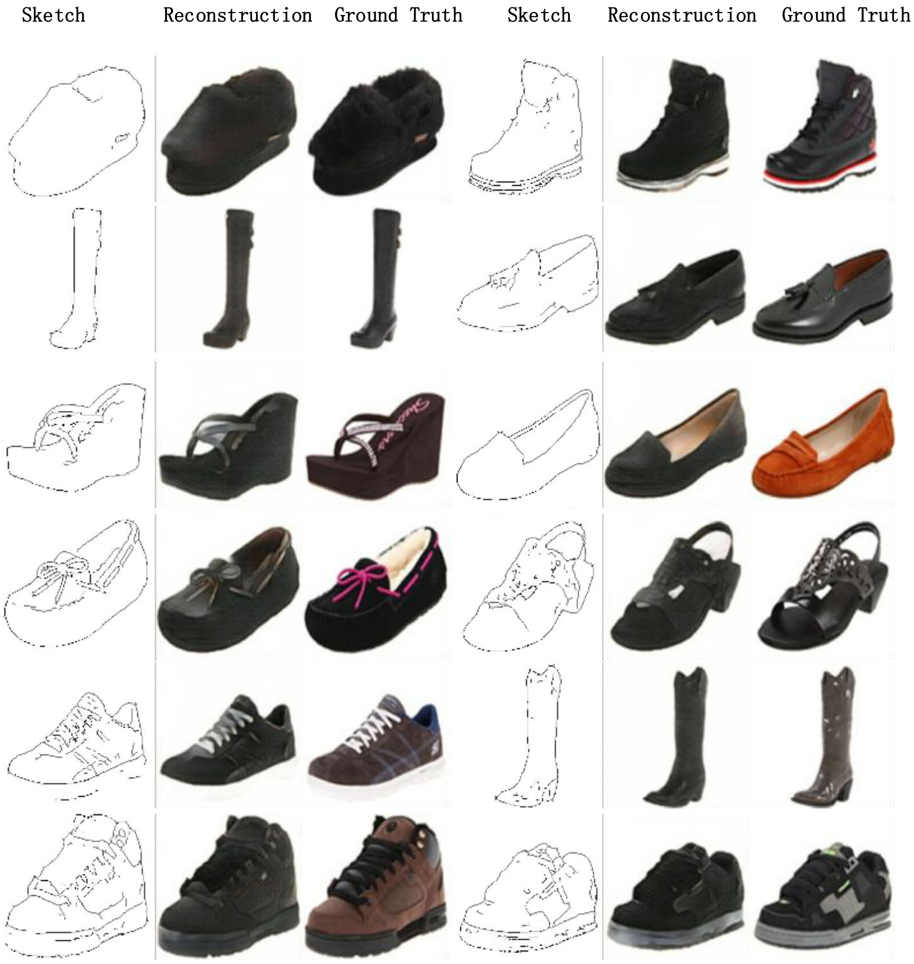

**Fig. 10** The sketch image to color image transfer results with our proposed method on the UT Zappos50K dataset. The first and fourth columns are input sketch, the second and fifth columns are reconstructed image, and the third and sixth columns are ground truth

using xDOG filter, which is more real-world scenarios. As shown in Fig. 10.

## 6 Conclusions

Based on the significance and current research of the image to image translation, in this paper, we propose a convolutional generative network with encoder-decoder architecture for face sketch to face image translation with multiple controllable facial attributes; we incorporate the adversarial loss, the facial attribute classification loss, and reconstruction loss into a unified network architecture. The results show that our proposed method not only can generate a realistic facial image in the map of the input sketch image but also can control the facial attributes change during face sketch to image translation.

However, the proposed network architecture also has shortcoming; the performance to manipulate some complex facial attributes is not satisfactory (e.g., eyeglasses, hat). The issue will be addressed in further research.

**Acknowledgements**
Thanks to my colleagues and classmates for their help in writing the paper; it is with their encouragement and guidance that I can finally complete this paper.

**Authors' contributions**
The first draft of the paper and experimental data are completed by MH, the collection of some experimental data, and the review of articles are completed by JG. Both authors read and approved the final manuscript.

**Authors' information**
Mingming Hu received M.S. degree from Beijing Jiaotong University, currently working in Institute of Software, Chinese Academy of Sciences. His research interests include artificial intelligence and image processing. Jingtao Guo is currently pursuing Ph.D degree from the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include artificial intelligence, image filling, and image processing.

**Funding**
The authors acknowledge the support from the National Key R&D Program "Research on Intersatellite Networking and Data Sharing Technology" (2016YFB0501104).

**Availability of data and materials**
Data and implementation codes for all experiments are based on python.

**Competing interests**
Not applicable

**Author details**
[1]Institute of Software, Chinese Academy of Sciences, Beijing 100190, China. [2]Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100000, China.

Received: 30 July 2019 Accepted: 31 December 2019
Published online: 13 January 2020

### References

1. Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 1005–1010. IEEE, 2005.
2. N. Wang, D. Tao, X. Gao, X. Li, J. Li, A comprehensive survey to face hallucination. International journal of computer vision **106**(1), 9–30 (2014)
3. H. Kazemi, M. Iranmanesh, A. Dabouei, et al., Facial attributes guided deep sketch-to-photo synthesis[C]// 2018 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE Computer Society (2018)
4. C. Peng, X. Gao, N. Wang, J. Li, Superpixel-based face sketch–photo synthesis. IEEE Transactions on Circuits and Systems for Video Technology **27**(2), 288–299 (2017)
5. M. Song, C. Chen, J. Bu, T. Sha, Image-based facial sketch-to-photo synthesis via online coupled dictionary learning. Information Sciences **193**, 233–246 (2012)
6. A. Brock, T. Lim, J.M. Ritchie, N. Weston, Neural photo editing with introspective adversarial networks. ArXiv preprint arXiv **1609**, 07093 (2016)
7. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of Wasserstein GANs. ArXiv preprint arXiv, 00028 (1704, 2017)
8. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, in *Advances in Neural Information Processing Systems (NIPS)*. Generative adversarial networks (2014)
9. A.B.L. Larsen, S.K. Sønderby, H. Larochelle, O. Winther, in *International Conference on Machine Learning (ICML)*. Autoencoding beyond pixels using a learned similarity metric (2016)
10. G. Perarnau, J. van de Weijer, B. Raducanu, J.M. Alvarez, in *Advances in Neural Information Processing Systems (NIPS) Workshops*. Invertible conditional GANs for image editing (2016)
11. M. Li, W. Zuo, D. Zhang, Deep identity-aware transfer of facial attributes. arXiv preprint arXiv **1610**, 05586 (2016)
12. W. Shen, R. Liu, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Learning residual images for face attribute manipulation (2017)
13. M.-Y. Liu, T. Breuel, J. Kautz, in *Advances in Neural Information Processing Systems (NIPS)*. Unsupervised image-to-image translation networks (2017)
14. S. Zhou, T. Xiao, Y. Yang, D. Feng, Q. He, W. He, in *British Machine Vision Conference (BMVC)*. Genegan: learning object transfiguration and attribute subspace from unpaired data (2017)
15. G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, M. Ranzato, in *Advances in Neural Information Processing Systems (NIPS)*. Fader networks: manipulating images by sliding attributes (2017)
16. T. Kim, B. Kim, M. Cha, J. Kim, Unsupervised visual attribute transfer with reconfigurable generative adversarial networks. arXiv preprint arXiv **1707**, 09798 (2017)
17. T. Xiao, J. Hong, J. Ma, in *International Conference on Learning Representations (ICLR) Workshops*. Dna-gan: learning disentangled representations from multi-attribute images (2018)
18. Sangkloy P, Lu J, Fang C, et al. Scribbler: controlling deep image synthesis with sketch and color[J]. 2016.
19. Zhang L, Lin L, Wu X, et al. End-to-end photo-sketch generation via fully convolutional representation learning [J]. 2015.
20. P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Image-to-image translation with conditional adversarial networks (2017)
21. M. Zhang, J. Li, N. Wang, et al., Compositional model-based sketch generator in facial entertainment[J]. IEEE Transactions on Cybernetics, 1–12 (2017)
22. M. Zhang, R. Wang, X. Gao, J. Li, D. Tao, "Dual-transfer face sketch–photo synthesis" IEEE Trans. Image Process. **28**(2), 642–657 (Feb. 2019)
23. M. Zhang, N. Wang, Y. Li, X. Gao, in *IEEE Transactions on Cybernetics*. Bionic face sketch generator
24. M. Zhang, N. Wang, Y. Li and X. Gao, Neural probabilistic graphical model for face sketch synthesis, in IEEE Transactions on Neural Networks and Learning Systems.
25. M. Zhang, Y. Li, N. Wang, Y. Chi and X. Gao, Cascaded face sketch synthesis under various illuminations, in IEEE Transactions on Image Processing.
26. N. Wang, X. Gao, J. Li, Random sampling for fast face sketch synthesis[J]. Pattern Recognition **76** (2017)
27. N. Wang, X. Gao, L. Sun, et al., Bayesian face sketch synthesis [J]. IEEE Transactions on Image Processing **PP**(99), 1 (2017)
28. N. Wang, X. Gao, L. Sun, J. Li, Anchored neighborhood index for face sketch synthesis. IEEE Transactions on Circuits and Systems for Video Technology **28**(9), 2154–2163 (Sept. 2018)
29. M. Zhu, J. Li, N. Wang, X. Gao, A deep collaborative framework for face photo–sketch synthesis. IEEE Transactions on Neural Networks and Learning Systems **30**(10), 3096–3108 (Oct. 2019)

30. C. Li and M. Wand. Combining Markov random fields and convolutional neural networks for image synthesis. CVPR, 2016.
31. L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks. CVPR (2016)
32. A. Dosovitskiy, T. Brox, Generating images with perceptual similarity metrics based on deep networks. arXiv preprint arXiv **1602**, 02644 (2016)
33. A. Efros, W.T. Freeman, in *SIGGRAPH, pages 341–346.ACM*. Image quilting for texture synthesis and transfer (2001)
34. M. Zhu, J. Li, N. Wang, et al., A deep collaborative framework for face photo-sketch synthesis [J]. IEEE Transactions on Neural Networks and Learning Systems, 1–13 (2019)
35. He Z, Zuo W, Kan M, et al. AttGAN: facial attribute editing by only changing what you want [J]. 2017.
36. Chang J, Scherer S. [IEEE ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - New Orleans, LA, USA (2017.3.5-2017.3.9)] 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - Learning representations of emotional speech with deep convolutional generative adversarial networks[J]. 2017:2746-2750.
37. Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of Wasserstein GANs[J]. 2017.
38. Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [J]. 2017.
39. M. Mirza, S. Osindero, Conditional generative adversarial nets. arXiv preprint arXiv **1411**, 1784 (2014)
40. A. Odena, Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv **1606**, 01583 (2016)
41. A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier GANs. arXiv preprint arXiv **1610**, 09585 (2016)
42. A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks [J]. Computer Science (2015)
43. Arjovsky M, Chintala S, Bottou, Léon. Wasserstein GAN[J]. 2017.
44. M. Li, W. Zuo, D. Zhang, *Convolutional network for attribute-driven and identity-preserving human face generation [J]* (2016)
45. G. Perarnau, J. van de Weijer, B. Raducanu, and J. M.Álvarez. Invertible conditional GANs for image editing. arXiv preprint arXiv:1611.06355, 2016.
46. Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild [J]. 2014.
47. H. WinnemoLler, J.E. Kyprianidis, S.C. Olsen, Xdog: an extended difference-of-Gaussians compendium including advanced image stylization. Computers & Graphics **36**(6), 740–753 (2012)

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.