Li HY, Dong WM, Hu BG. Facial image attributes transformation via conditional recycle generative adversarial networks. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 33(3): 511–521 May 2018. DOI 10.1007/s11390-018-1835-2

Facial Image Attributes Transformation via Conditional Recycle Generative Adversarial Networks

Huai-Yu Li[1,2], Wei-Ming Dong[1,*], *Member*, *CCF*, *ACM*, *IEEE*, and
Bao-Gang Hu[1], *Senior Member*, *IEEE*, *Member*, *CCF*

[1] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[2] *University of Chinese Academy of Sciences, Beijing 100049, China*

E-mail: huaiyu.li@nlpr.ia.ac.cn; weiming.dong@ia.ac.cn; hubg@nlpr.ia.ac.cn

Received December 25, 2017; revised March 20, 2018.

Abstract This study introduces a novel conditional recycle generative adversarial network for facial attribute transformation, which can transform high-level semantic face attributes without changing the identity. In our approach, we input a source facial image to the conditional generator with target attribute condition to generate a face with the target attribute. Then we recycle the generated face back to the same conditional generator with source attribute condition. A face which should be similar to that of the source face in personal identity and facial attributes is generated. Hence, we introduce a recycle reconstruction loss to enforce the final generated facial image and the source facial image to be identical. Evaluations on the CelebA dataset demonstrate the effectiveness of our approach. Qualitative results show that our approach can learn and generate high-quality identity-preserving facial images with specified attributes.

Keywords generative adversarial network, image editing, facial attributes transformation

1 Introduction

Social media has become popular platforms for the public to share personal photos. People often apply facial editing or facial beautification operations to make photos visually attractive. Facial image manipulation has become an active research area in computer graphics and computer vision. Users can interactively transform their appearances by stylization[1], compositing[2], virtual makeup[3-4], etc. In addition, modifying facial images can also help improve the accuracy of face recognition systems for a few hard facial images[5-7].

Although existing facial image editing methods can effectively generate virtual appearances for people, transforming facial features using high-level semantic attributes remains challenging. The transformation of facial attributes can help people visualize what they may look like with different adornments, expressions, hairstyles, genders and so on. Traditional ways to edit facial image attributes and acquire realistic results usually require skilled users with image editing software such as Adobe Photoshop, or operate by manipulating patches of existing images[8]. Recently, benefited from the development of deep learning techniques, many generative models[9-11] are able to synthesize promising realistic images. Among these models, generative adversarial networks (GANs)[12] are one of the most efficient approaches for image generation.

In this paper, we present a novel framework to automatically transform the face in a given image to a new appearance with specified attributes and identical personal identity. We design a conditional recycle GAN (CRGAN) for the identity-preserving facial image transformation to generate realistic and high-quality results.

Regular Paper

Special Section of CVM 2018

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61672520, 61573348, 61620106003, and 61720106006, the Beijing Natural Science Foundation of China under Grant No. 4162056, the National Key Technology Research and Development Program of China under Grant No. 2015BAH53F02, and the CASIA-Tencent YouTu Jointly Research Project. The Titan X used for this research was donated by the NVIDIA Corporation.

*Corresponding Author

In our framework, we first use a conditional generator to transform a given face to a face with target attributes and use a discriminator to judge whether the generated face is real and predict its attributes. We then recycle the generated face back to the same conditional generator and transform it back to a face with original attributes. We introduce recycle reconstruction loss to constrain the generated image in the recycle phase to maintain personal facial identity. Our method is a data-driven approach that directly learns facial transformation from face datasets with attribute annotations. After training, the generator learns to transform a given face image with specified attributes. Hence, given a facial image and the target facial attributes, the learned generator can transform the given face into a face with target attributes without changing face identity. As shown in Fig.1, when we specify a Black Hair attribute for a black girl face, we modify her hair color into blond; when we specify an Eyeglasses attribute, she wears an eyeglass; even when we want her face to grow a mustache, we only need to simply specify a Mustache attribute. Therefore, we can explicitly manipulate the attributes of a real facial image by learning CRGANs, and the model may also be generally applied to other image attributes manipulation tasks.

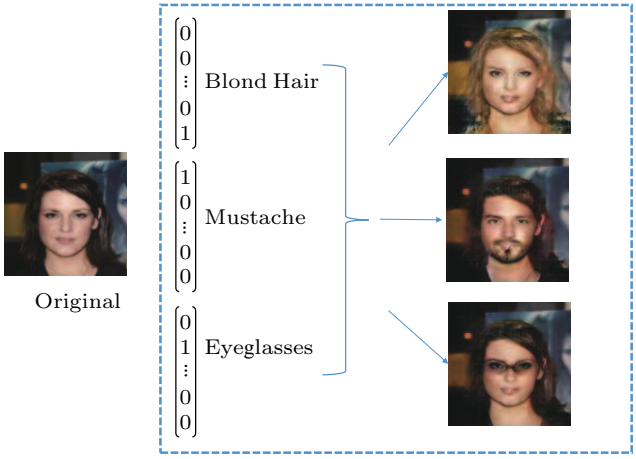


Fig.1. Example of conditional facial image generation from visual attributes.

The contributions of this paper are summarized as follows.

• We propose a novel conditional GAN (CGAN) learning architecture named CRGAN for learning identity-preserving facial image transformation.

• We propose direct learning from attribute-annotated facial image datasets without any landmarks to transform multiple attributes for a given face.

The rest of this paper is organized as follows. Section 2 describes related work on facial image manipulation. Section 3 provides a detailed description of the CRGAN architecture. Section 4 presents a few experiments to validate the effectiveness of the proposed method in facial image transformation. Finally, Section 5 provides the CRGAN result analysis and discusses the limitations and future directions.

2 Related Work

Different approaches are available for face image attribute transformation. Among computational photography methods, Kemelmacher-Shlizerman[2] presented a system that enables automatic synthesis of unlimited numbers of appearances. The system uses photo and text queries as inputs. The text descriptions are used to retrieve related photos via a web image engine. Retrieved photo features are then computed and matched to the input photo. The input face is finally blended into the highest ranked candidates. Although this method can produce many impressive results, the limitation lies in retrieved photo quality. The synthesized photos may occasionally look unnatural and unreal.

Image attributes transformation techniques have leveraged deep convolutional neural networks in recent years. Upchurch *et al.*[13] proposed a data-driven approach named deep feature interpolation (DFI) for automatic transformation of high-resolution image. This method applies a simple linear interpolation between the source image and target attribute features to obtain attribute features. A neural network is then trained to generate an image with the attribute features. The features are extracted from the pretrained network that distinguishes between images with and without those attributes. Although the method can generate high-quality transformed images, the apparent limitations lie in the aligned image requirement and the similarity of target image to sample images for the pretrained network.

Peraranau *et al.*[14] proposed an invertible conditional GAN (IcGAN) that learns real image mapping into a latent space and a conditional representation. The encoded image latent space and conditional representation are used to reconstruct and modify real images of face conditioning on arbitrary attributes. However, the approach separates the learning procedure into several steps. It needs to pretrain the encoder and then prepares the dataset to train generators. Furthermore, this approach only applies to low-resolution facial images in 64×64.

Yin *et al.*[15] proposed a semi-latent GAN that learns the relationship between user-defined and latent attributes and between attributes and images to generate and modify facial images from high-level semantic attributes. The learning framework is novel, but the results present low quality and are blurry. Liao *et al.*[16] transfered visual attributes across images that may have different appearances but exhibit perceptually similar semantic structures. This approach can transfer an image to different domains, whereas our approach focuses on adding or removing specific attributes. Lu *et al.*[17] extended CycleGAN[18] to conditional CycleGAN, in which the mapping from X to Y is subjected to attribute condition. The framework comprises two pairs of generators and discriminators: one is learning to map from high-resolution images to low-resolution images, and the other is learning to map from low-resolution images with attribute condition to high-resolution target images. An additional face verification loss is introduced into the training process to preserve facial image identity. This method can generate high-resolution identity-preserving face images with transformed attributes; however, unrealistic results are occasionally produced. Unlike this approach, our framework only contains one pair of CGAN and learns to transform the transformed facial image back to the original one.

Recently, Choi *et al.*[19] proposed StarGAN that can learn and perform image-to-image translations for multiple domains using a single model. Their approach is similar to ours which contains one pair of the generator and discriminator and uses reconstruction loss to train the generator. The difference lies on how to introduce attribute conditions into the generator. StarGAN concatenates the attribute vectors to the input facial images of the generator, while our method concatenates the attribute conditions to the bottleneck features. StarGAN also allows to simultaneously train multiple datasets with different domains within a single network.

Our approach ameliorates the limitations of previously mentioned approaches. Moreover, this approach does not require specifically pretrained network, or multi-scale transformation from low-resolution images to generate high-resolution images. Our CRGAN framework is simple and can transform high-quality images that are realistic and natural.

3 Method

We describe our approach for facial image attribute transformation with CRGAN in detail in this section. First, we introduce several GAN architectures related to our approach. Second, we describe the CRGAN framework. Finally, we summarize the full objective of CRGAN.

3.1 Related GAN Architectures

A GAN[12] comprises two neural networks, namely, a generator G and a discriminator D, which are iteratively trained in a two-player minimax game manner to learn to generate samples that are similar to real data samples. GAN is optimized using the adversarial loss $\mathcal{L}(G, D)$ defined as

$$\begin{aligned} &\min_G \max_D \mathcal{L}(G, D) \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}(\log D(\boldsymbol{x})) + \\ &\quad \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}(\log(1 - D(G(\boldsymbol{z})))), \end{aligned} \tag{1}$$

where $p_{\text{data}}(\boldsymbol{x})$ is the distribution of real data and $p_z(\boldsymbol{z})$ is a prior distribution (usually a standard Gaussian distribution). The parameters of generator G and discriminator D are iteratively updated in training. For image generation, it draws a sample $\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})$ from the prior distribution and then transforms it through the generator to obtain images $G(\boldsymbol{z})$.

CGAN[20] is a conditional version of GANs, which can be constructed by feeding additional information $\boldsymbol{c}$ both the generator and discriminator condition on. In this way, the generated samples can be directly controlled by a specified condition. The objective function of CGAN is formulated as follows:

$$\begin{aligned} &\min_G \max_D \mathcal{L}(G, D) \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}(\log D(\boldsymbol{x}|\boldsymbol{y})) + \\ &\quad \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}(\log(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))), \end{aligned}$$

where $\boldsymbol{y}$ denotes the additional information conditions for the generator and the discriminator. Condition $\boldsymbol{y}$ could be any type of auxiliary information, such as class labels or features from other modalities. Other notations are similar to those in (1).

Auxiliary classifier GAN (AC-GAN)[21] is another variant of the GAN architecture. Apart from noise $\boldsymbol{z}$, every generated sample in AC-GAN has a corresponding class label denoted as $\boldsymbol{c} \sim p_c(\boldsymbol{c})$. The generator is similar to that in CGAN which uses $\boldsymbol{c}$ and $\boldsymbol{z}$ to generate images $X_{\text{fake}} = G(\boldsymbol{c}, \boldsymbol{z})$. However, the discriminator produces a probability distribution over images $P(S|X)$ and a probability distribution over class labels $P(C|X)$. The objective function comprises the log-likelihood of

the correct source image $\mathcal{L}_S$ and the log-likelihood of the correct class $\mathcal{L}_C$ as follows:

$$\begin{aligned}\mathcal{L}_S &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}(\log P(S=\text{real}|\boldsymbol{x}))+\\ &\quad \mathbb{E}_{\boldsymbol{x}\sim p_{\text{fake}}(\boldsymbol{x})}(\log P(S=\text{fake}|\boldsymbol{x})),\\ \mathcal{L}_C &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}(\log P(C=c|\boldsymbol{x}))+\\ &\quad \mathbb{E}_{\boldsymbol{x}\sim p_{\text{fake}}(\boldsymbol{x})}(\log P(C=c|\boldsymbol{x})).\end{aligned}$$

The discriminator is trained to maximize $\mathcal{L}_S+\mathcal{L}_C$, whereas the generator is trained to maximize $\mathcal{L}_C-\mathcal{L}_S$.

The goal of CycleGAN[18] is to learn mapping functions between two domains X and Y given training samples $\{x_i\}_{i=1}^N \in X$ and $\{y_j\}_{j=1}^M \in Y$, respectively. The translations $X\to Y$ and $Y\to X$ are simultaneously learned with two generators and two discriminators. Unlike pix2pix[22] which requires paired training data, CycleGAN only needs unpaired training data. Hence, cycle consistency loss is proposed to enforce forward-backward consistency which ensures that the image translation cycle for each image x from domain X can translate x back to the original image. With cycle consistency, the objective function of CycleGAN is defined as:

$$\begin{aligned}&\mathcal{L}(G,F,D_X,D_Y)\\ &=\mathcal{L}_{\text{GAN}}(G,D_Y,X,Y)+\mathcal{L}_{\text{GAN}}(F,D_X,Y,X)+\\ &\quad \lambda\mathcal{L}_{\text{cyc}}(G,F),\end{aligned}$$

where $\mathcal{L}_{\text{cyc}}(G,F)=\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}(\boldsymbol{x})\parallel F(G(\boldsymbol{x}))-\boldsymbol{x}\parallel_1+\mathbb{E}_{\boldsymbol{y}\sim p_{\text{data}}}(\boldsymbol{y})\parallel G(F(\boldsymbol{y}))-\boldsymbol{y}\parallel_1$ is the cycle consistency loss, and $\mathcal{L}_{\text{GAN}}(G,D_Y,X,Y)$ and $\mathcal{L}_{\text{GAN}}(F,D_X,Y,X)$ are the adversarial losses for translations $X\to Y$ and $Y\to X$, respectively.

Motivated by CycleGAN, we combine CGAN and ACGAN in our work to construct an adversarial training architecture that can directly learn from facial image datasets and transform input facial images with specified conditional semantic attributes without changing the person's face identity.

3.2 Conditional Recycle Generative Adversarial Network

Our goal is to learn a generation network of high-resolution facial images that can transform a given facial image by specifying high-level semantic attributes with unchanged facial identity. We accordingly design a novel generative adversarial learning network architecture to realize this goal, as illustrated in Fig.2. The model comprises two transformation phases but with one conditional generator and one adversarial discriminator, because the parameters of the generator and discriminator in the two phases are shared.

We regard the facial images with target attributes as real samples and the generated facial images with target attributes as fake samples to train the attribute conditional GAN network. Thus, in the first transformation phase, the conditional generator attempts to transform the original facial image $\boldsymbol{x}$ with condition $\boldsymbol{c_y}$

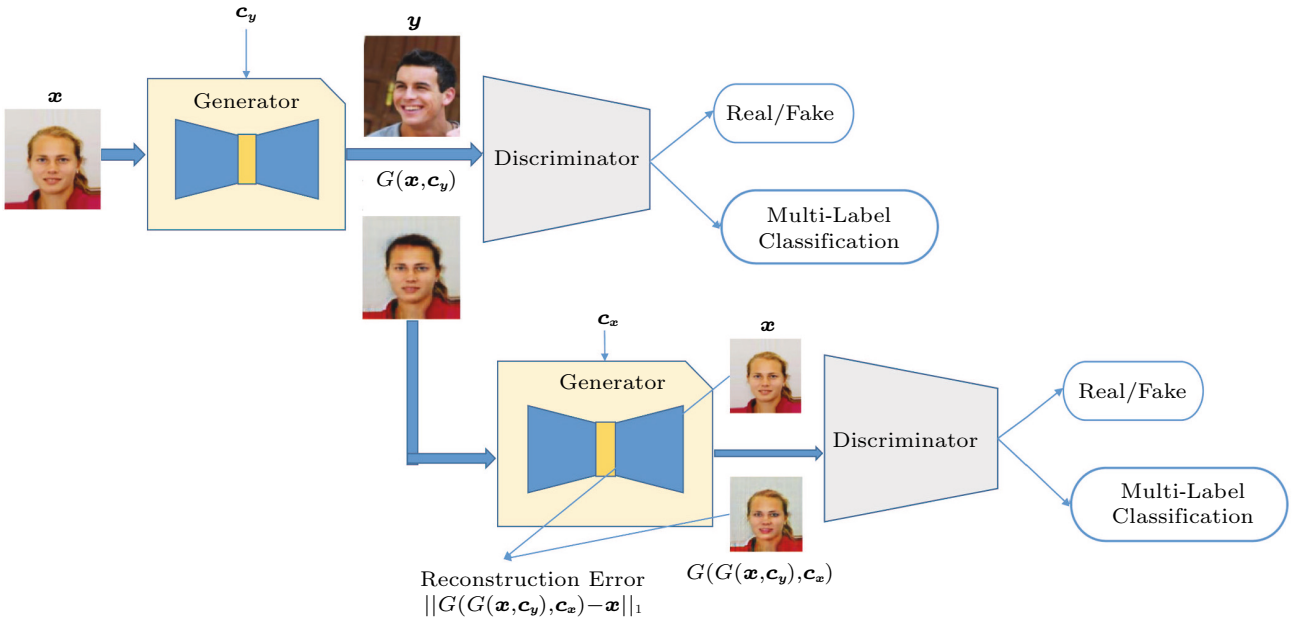


Fig.2. Overview of our CRGAN architecture for facial image attribute transformation. The generators and discriminators in two phases share the same weights.

to generate facial image $G(\boldsymbol{x}, \boldsymbol{c_y})$ with $\boldsymbol{c_y}$ attributes. The discriminator D aims at distinguishing between generated image $G(\boldsymbol{x}, \boldsymbol{c_y})$ with attribute $\boldsymbol{c_y}$ and real image $\boldsymbol{y}$ with attribute $\boldsymbol{c_y}$. Simultaneously, we use a multi-label classification loss to predict the attributes contained in the facial images to explicitly preserve facial attributes. We define the conditional auxiliary classifier adversarial loss for this phase as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}_1} &= \mathbb{E}_{\boldsymbol{y},\boldsymbol{c_y}\sim p_{\mathrm{real}}(\boldsymbol{y},\boldsymbol{c_y})}\log(D(\boldsymbol{y})) + \\ &\quad \mathbb{E}_{\boldsymbol{x},\boldsymbol{c_y}\sim p_{\mathrm{fake}}(\boldsymbol{x},\boldsymbol{c_y})}(\log(1 - D(G(\boldsymbol{x}, \boldsymbol{c_y})))), \\ \mathcal{L}_{\mathrm{C}_1} &= \mathbb{E}_{\boldsymbol{y},\boldsymbol{c_y}\sim p_{\mathrm{real}}(\boldsymbol{y},\boldsymbol{c_y})}H(P(\boldsymbol{y}), \boldsymbol{c_y}) + \\ &\quad \mathbb{E}_{\boldsymbol{x},\boldsymbol{c_y}\sim p_{\mathrm{fake}}(\boldsymbol{x},\boldsymbol{c_y})}H(P(G(\boldsymbol{x}, \boldsymbol{c_y})), \boldsymbol{c_y}), \qquad (2)\end{aligned}$$

where $\mathcal{L}_{\mathrm{GAN}_1}$ denotes the generative adversarial loss. $P(\boldsymbol{y})$ and $P(G(\boldsymbol{x}, \boldsymbol{c_y}))$ denote the probability vectors for each attribute of real and generated facial images, respectively. $H(P(\boldsymbol{y}), \boldsymbol{c_y})$ is the cross entropy loss between the probability of real facial image attributes and target labels, and $H(P(G(\boldsymbol{x}, \boldsymbol{c_y}), \boldsymbol{c_y})$ is the cross entropy loss between the probability of generated facial image attributes and target labels. A facial image usually contains several attributes; thus, $\mathcal{L}_{\mathrm{C}_1}$ denotes multi-label classification loss. The training of the discriminator to minimize $\mathcal{L}_{\mathrm{C}_1}$ guides the discriminator in discerning attribute features. The training of the generator to minimize $\mathcal{L}_{\mathrm{C}_1}$ guides the generator to produce images with target attributes.

The first phase of adversarial training can, theoretically, learn mapping G that produces facial images identically distributed as target attribute facial image domains. A network with a sufficiently large capacity can map the same set of input images to any random permutation of images in the target domain[18]. However, training in practice with only the first phase losses (2) usually produces identical facial images with inconspicuous attributes modification. Therefore, the first phase losses alone cannot achieve our goal. We introduce the recycle generative adversarial learning phase to transform facial image attributes with evident modifications and without modifying personal identity. In this phase, we recycle conditionally generated images $G(\boldsymbol{x}, \boldsymbol{c_y})$ as inputs back to the same adversarial training network. However, we use $\boldsymbol{c_x}$, which is the attributes of facial image $\boldsymbol{x}$, as the condition. The generated facial images are then denoted as $G(G(\boldsymbol{x}, \boldsymbol{c_y}), \boldsymbol{c_x})$. We enforce the facial images through the two transformation phases should be recycle-consistent, which means $\boldsymbol{x} \to G(\boldsymbol{x}, \boldsymbol{c_y}) \to G(G(\boldsymbol{x}, \boldsymbol{c_y}), \boldsymbol{c_x}) \approx \boldsymbol{x}$. Hence, we define the losses for this phase as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}_2} &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{c_x}\sim p_{\mathrm{real}}(\boldsymbol{x},\boldsymbol{c_x})}\log(D(\boldsymbol{x})) + \\ &\quad \mathbb{E}_{\boldsymbol{x},\boldsymbol{c_y},\boldsymbol{c_x}\sim p_{\mathrm{fake}}(\boldsymbol{x},\boldsymbol{c_y},\boldsymbol{c_x})} \\ &\quad (\log(1 - D(G(G(\boldsymbol{x}, \boldsymbol{c_y}), \boldsymbol{c_x})))), \\ \mathcal{L}_{\mathrm{C}_2} &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{c_x}\sim p_{\mathrm{real}}(\boldsymbol{x},\boldsymbol{c_x})}H(P(\boldsymbol{x}), \boldsymbol{c_x}) + \\ &\quad \mathbb{E}_{\boldsymbol{x},\boldsymbol{c_y},\boldsymbol{c_x}\sim p_{\mathrm{fake}}(\boldsymbol{x},\boldsymbol{c_y},\boldsymbol{c_x})} \\ &\quad H(P(G(G(\boldsymbol{x}, \boldsymbol{c_y}), \boldsymbol{c_x})), \boldsymbol{c_x}), \\ \mathcal{L}_{\mathrm{recyc}} &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{c_y},\boldsymbol{c_x}\sim p_{\mathrm{data}(\boldsymbol{x},\boldsymbol{c_y},\boldsymbol{c_x})} \\ &\quad (\parallel G(G(\boldsymbol{x}, \boldsymbol{c_y}), \boldsymbol{c_x}) - \boldsymbol{x} \parallel_1),\end{aligned}$$

where $\mathcal{L}_{\mathrm{GAN}_2}$ and $\mathcal{L}_{\mathrm{C}_2}$ are the adversarial and multi-label classification losses, respectively, which are similar to those in the first phase. $\mathcal{L}_{\mathrm{recyc}}$ is the recycle reconstruction loss. We use $L1$ distance to measure the difference between the generated facial images through the two phases with the original image. The minimization of this reconstruction loss aims to maintain the identity of transformed facial images.

We add an anisotropic total variation loss[23] to slightly smooth the generated facial images. The loss is defined for the transformed facial image $G(\boldsymbol{x}, \boldsymbol{c_y})$ and the recycled facial image $G(G(\boldsymbol{x}, \boldsymbol{c_y}), \boldsymbol{c_x})$. We define $z = [z_{ij}] = G(\boldsymbol{x})$ as

$$\mathcal{L}_{TV}(z) = \sum_{i,j}((z_{i,j+1} - z_{i,j})^2 + (z_{i+1,j} - z_{i,j})^2)^{\frac{B}{2}},$$

where we set $B = 2$ in all our experiments.

3.3 Full Objective

In summary, the full objective is formulated as:

$$\begin{aligned}\mathcal{L}(G, D) &= \mathcal{L}_{\mathrm{GAN}_1} + \mathcal{L}_{\mathrm{GAN}_2} + \mathcal{L}_{\mathrm{C}_1} + \mathcal{L}_{\mathrm{C}_2} + \\ &\quad \lambda\mathcal{L}_{\mathrm{recyc}} + \beta\mathcal{L}_{\mathrm{TV}},\end{aligned}$$

where λ and β control the relative importance of different objectives. We aim to solve:

$$G^*, D^* = \arg\min_G \max_D \mathcal{L}(G, D),$$

which is a minmax optimization problem that requires careful optimization. The full objective is trained in the standard adversarial training scheme[12]. Notably, the parameters of the generator and the discriminator are shared in the two phases. Therefore, the gradients computed from their corresponding objectives are accumulated. We first update the discriminator parameters via maximizing

$$\mathcal{L}_{\mathrm{GAN}_1} + \mathcal{L}_{\mathrm{GAN}_2} + \mathcal{L}_{\mathrm{C}_1} + \mathcal{L}_{\mathrm{C}_2}.$$

We second maintain the discriminator fixed and update the generator parameters via minimizing

$$\mathcal{L}_{\mathrm{GAN}_1} + \mathcal{L}_{\mathrm{GAN}_2} - \mathcal{L}_{\mathrm{C}_1} - \mathcal{L}_{\mathrm{C}_2} + \lambda\mathcal{L}_{\mathrm{recyc}} + \beta\mathcal{L}_{\mathrm{TV}}.$$

Finally, we train the model by iterating this alternative process.

Moreover, the generator and the discriminator of our CRGAN have a special internal neural network structure, which will be elaborated in Section 4.

4 Implementation and Experiments

We demonstrate in this section the efficacy of our method in controlling the transformation task of facial image semantic attributes. First, we introduce the experimental datasets. Second, we describe the implementation of our architecture and training details. Third, we demonstrate both the single and multiple attributes facial image transformation results. Finally, we qualitatively compare the results with those of IcGAN and StarGAN and discuss the limitations of our model.

4.1 Experimental Dataset

We conduct the experiments on the CelebA dataset[24] which contains 202 599 facial images of celebrities with shape 178×218, and 40 different attribute labels, where each label is a binary value. We use facial images with 17 selected attributes for our experiments because the number of facial images with a few attributes is relatively small while a few attributes are not visually evident. The selected facial attributes are as follows: Bald, Bangs, Black Hair, Blond Hair, Gray Hair, Brown Hair, Mustache, Pale Skin, Smiling, Eyeglasses, Gender, Bushy Eyebrows, Attractive, Young, Heavy Makeup, Wearing Hat, and Narrow Eyes. We use a standard training and evaluation dataset split on aligned and cropped version and select the facial images with at least one of the 17 attributes. Hence, 182 111 images for training and 19 869 images for evaluation are selected.

4.2 Implementation and Training Details

We adopt the architecture of the network in [22] to build our CRGAN framework. We use U-Net[25] architecture as our conditional generator, which is an encoder-decoder structure that allows low-level information to shortcut across the network, thereby leading to improved results. We illustrate the general shape of the conditional U-Net in Fig.3.

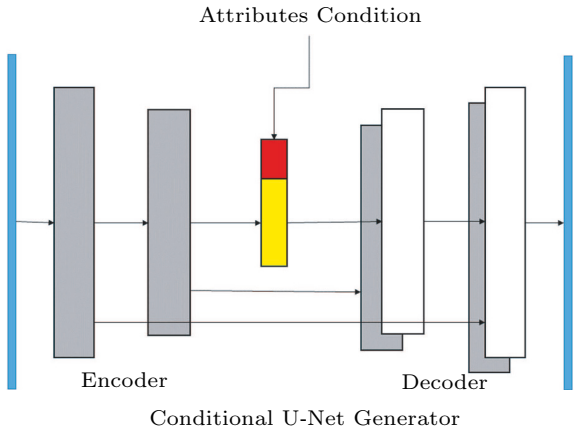


Fig.3. Illustration of conditional U-Net generator general structure. The black arrow represents skip connections between the encoder and the decoder. The overlap in the decoder indicates the concatenation of features. The yellow block denotes the latent representation, and the red block is the attribute condition vector.

We let C_k denote a Convolution-BatchNorm[26]-Leaky_ReLu[27] layer in the encoder with k filters. The slope of Leaky-ReLu is 0.2. Convolutions use a kernel of size 5×5, with a stride of 2 and "SAME" padding; thus, each layer of the encoder divides the size of its input by 2. The input images have a shape of $128 \times 128 \times 3$. The encoder comprises the following seven layers:

$$C_{64} - C_{128} - C_{256} - C_{512} - C_{512} - C_{512} - C_{512}.$$

Consequently, the latent representation of an image is a feature map with a size $512 \times 1 \times 1$. We then use one-hot vectors to encode attributes of facial images and concatenate them to the latent representations. DC_k is denoted as a Deconvolution[28]-BatchNorm[26]-ReLu layer in the decoder with k filters. The U-Net structure concatenates the encoder features to the corresponding decoder layers because skip connections exist between the encoder and the decoder. Hence, the decoder comprises the following seven layers:

$$DC_{64\times2} - DC_{128\times2} - DC_{256\times2} - DC_{512\times2} - DC_{512\times2} - C_{512\times2} - C_{512\times2}.$$

We use fc_n to denote the fully-connected layer with n hidden units. The discriminator is a five-Convolution-BatchNorm-Leaky_ReLu layer neural network, followed by two fully-connected layers branches. One branch is a fully-connected layer with an output size of 1, which is used to discriminate a real/fake image. The other branch is a fully-connected layer with an output size of 17, which is used to predict the probability for each attribute. The discriminator is described as follows:

$$C_{64} - C_{128} - C_{256} - C_{374} - C_{512} - fc_1,$$

$$C_{64} - C_{128} - C_{256} - C_{374} - C_{512} - fc_{17},$$

where the convolutional layers share the same weights.

Our method is implemented with a deep learning framework, namely, Tensorflow[29]. For preprocessing, we resize the images to 128×128, which is also the resolution used in all results presented in this paper. Image values are normalized to $[-1, 1]$. We do not perform any other data augmentation on training images. All models are trained with stochastic gradient descent algorithm with Adam[30] solver by using a learning rate of 0.000 2, $\beta_1 = 0.5$, and a batch size of 96. We set the hyperparamters of recycle reconstruction loss $\lambda = 0.000\,1$ and the total variation loss $\beta = 0.000\,1$. The training procedure follows the training scheme described in Subsection 3.3. The experiment takes approximately 10 hours of training on a single Pascal Titan X GPU and consumes around 8 Gb (Gigabit) GPU memory. All results presented in this paper or used for evaluation are taken from a validation set.

4.3 Single Attribute Transformation

We evaluate the performance of facial image transformation by changing one attribute, as shown in the results in Fig.4. Notably, a few inherent difficulties exist among some attributes. For example, the facial attributes, Bald and Wearing Hat, and Bald and Hair Style, conflict with each other. A few attributes may not generate distinguishable changes, such as Attractive, Young, Heavy Makeup. Hence, we show evident transformed facial results with one attribute modified in their corresponding attribute conditions in Fig.4. Clearly, the transformed results conditioned on attributes, such as Bangs, Black Hair, Blond Hair, Mustache, Smiling, Narrow Eyes, are convincing. For each row, the facial identity is well preserved. The target attributes can be either added to the face or removed from the face.

4.4 Multi-Attribute Transformation

We present the capability of our model to simultaneously modify multiple attributes in Fig.5. Fig.5(a) presents the original facial images. Fig.5(b) jointly modifies Blond Hair, Bangs, and Eyeglasses. Fig.5(c) jointly modifies Black Hair, Smiling, and Narrow Eyes. Fig.5(d) jointly modifies Black Hair, Pale Skin, and Bushy Eyebrows. Fig.5(e) jointly modifies Blond Hair, Mustache, and Gender. Bangs are modified in the same color with Blond Hair. A female face is transformed into a male-like face with mustache and blond hair. The transformed face can still be recognized as the same person as before. Even in these difficult settings, our model can still generate convincing identity-preserving facial images with multiple modified attribute conditions.

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n)

Fig.4. CRGAN results of one attribute modification on a given facial image. Each row is the same face with different modified attributes. (a) Original facial image. The followings are facial images with modified attributes: (b) Bald, (c) Bangs, (d) Black Hair, (e) Blond Hair, (f) Gray Hair, (g) Brown Hair, (h) Mustache, (i) Pale Skin, (j) Smiling, (k) Eyeglasses, (l) Gender, (m) Bushy Eyebrows, and (n) Narrow Eyes.

Fig.5. CRGAN results of simultaneous multiple attribute modification on given face images. (a) Original facial images. (b) Jointly modifying Blond Hair, Bangs, and Eyeglasses. (c) Jointly modifying Black Hair, Smiling, and Narrow Eyes. (d) Jointly modifying Black Hair, Pale Skin, and Bushy Eyebrows. (e) Jointly modifying Blond Hair, Mustache, and Gender.

4.5 Comparison

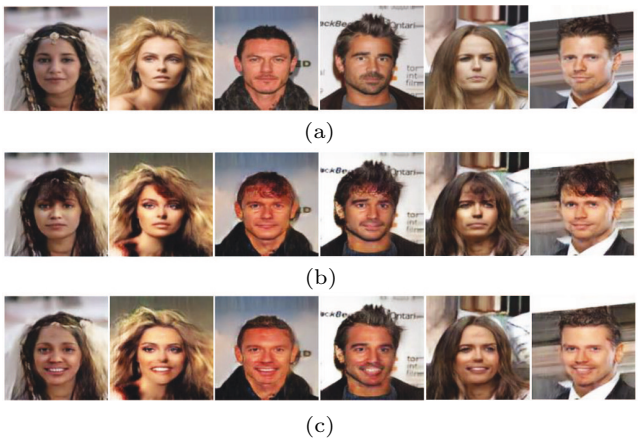
We qualitatively compare our results with those of IcGAN and StarGAN to verify the efficacy of our architecture, as illustrated in Fig.6. Evidently, our results have three advantages over IcGANs. First, the transformed facial images in our approach have a higher resolution of 128×128 than those in IcGANs (that is, 64×64). Central crop operation on the faces is also unnecessary. Second, our results have better visual quality and salient attributes modification. Third, our results have never changed facial identity. When compared with StarGAN which is the state-of-the-art approach, our results also exhibit competitive visual performance. We both have distinct facial attributes transformation results.

5 Analysis and Discussion

We demonstrate the diversity and adaption of conditional attribute results transformed by our CRGAN

Fig.6. Qualitative results comparison of CRGAN, IcGAN and StarGAN. (a) Original facial images. The followings are facial images with transformed attributes: (b) Bald, (c) Bangs, (d) Black Hair, (e) Blonde Hair, (f) Eyeglasses, (g) Heavy Makeup, (h) Gender, (i) Pale Skin, and (j) Smiling.

in Fig.7. We use Bangs and Smiling attributes as examples.



(a) (b) (c)

Fig.7. Diversity of facial attribute transformation. (a) Original facial images. (b) Transforming Bangs attribute. (c) Transforming Smiling attribute.

By adding the Bangs attribute to the input faces, we can see that for different hair styles and hair colors, the conditional generator can generate different kinds of bangs to suit different faces. The generated bangs are also influenced by the original face styles. For different female and male faces, although these different kinds of bangs are not explicitly annotated in the dataset, CRGAN can generate straight bangs, angled bangs, curled bangs, and tilted bangs. By adding Smiling attribute to the input faces, CRGAN not only simply opens the mouth of the face, but also controls muscles in the face and modifies the shape of the eyes. Therefore, the transformed facial images are natural and realistic.

Exploring the approach of CRGAN in learning facial image transformation with target attributes in an identity-preserving way is interesting. Although the facial images are all aligned in the datasets, numerous irregular faces are still blurry, inclined and side faces. Moreover, the faces in the training dataset are not segmented into semantic components, such as hair and jaw; we only use weak attribute annotations. From our perspective, we consider that CRGAN can gradually learn the relationships between weak attribute annotations and their corresponding regions. The textures of the attribute corresponding regions are controlled by the specified attribute conditions.

Although our method can achieve many compelling results in various cases, the results are still not uniformly consistent. In previous work[18,22-23,31-32], variant GANs can successfully transform a given image to a target style domain, but the transformation mostly involves image color and texture changes. Succeeding in making meaningfully geometric deformation for an image object remains difficult. In our CRGAN model, the learned conditional generator can make a few small appearance transfigurations such as Smiling, Bushy Eyebrows, Bangs, and Narrow Eyes with the help of weakly supervised attribute labels. However, successfully transforming the attributes, such as Wearing Hat and Eyeglasses, is still difficult because these attributes require a considerable extent of deformation. As shown in the failure cases in Fig.8, the failure case in adding eyeglass usually leads to the addition of two black blocks on the eyes or two dark circles around the eyes. Meanwhile, the transformed results are unsatisfactory for the addition of wearing a hat which requires a considerable large extent deformation.

(a) (b) (c)

Fig.8. Failure cases for geometric deformation attribute transformation. (a) Original facial image. (b) Transforming Eyeglasses attribute. (c) Transforming Wearing Hat attribute.

6 Conclusions

In this paper, we proposed CRGAN for learning identity-preserving high-quality facial image transformation task. Our main contribution lies in utilizing conditional generator and recycle reconstruction loss to build a GAN learning framework that can directly learn from facial images with multiple attributes. Our framework is clearer and simpler than other relevant methods[14-15,17] and can generate single attribute and multi-attributes transformation. Furthermore, our

method is a general method that can be easily extended to other attribute transformation tasks. In the future, we will improve the CRGAN capability in geometric deformation for image objects and apply it to other general image editing problems.

References

[1] Selim A, Elgharib M, Doyle L. Painting style transfer for head portraits using convolutional neural networks. *ACM Trans. Graph.*, July 2016, 35(4): 129:1–129:18.

[2] Kemelmacher-Shlizerman I. Transfiguring portraits. *ACM Trans. Graph.*, 2016, 35(4): 94:1–94:8.

[3] Li C, Zhou K, Lin S. Simulating makeup through physics-based manipulation of intrinsic image layers. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2015, pp.4621–4629.

[4] Liu S, Ou X, Qian R, Wang W, Cao X. Makeup like a superstar: Deep localized makeup transfer network. In *Proc. the 25th International Joint Conference on Artificial Intelligence*, July 2016, pp.2568–2575.

[5] Tran L, Yin X, Liu X. Disentangled representation learning gan for pose-invariant face recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2017, pp.1283–1292.

[6] Chen C, Dantcheva A, Ross A. Automatic facial makeup detection with application in face recognition. In *Proc. International Conference on Biometrics*, February 2013.

[7] Yao N M, Chen H, Guo Q P, Wang H A. Non-frontal facial expression recognition using a depth-patch based deep neural network. *Journal of Computer Science and Technology*, 2017, 32(6): 1172–1185.

[8] Mohammed U, Prince S, Kautz J. Visio-lization: Generating novel facial images. *ACM Trans. Graph.*, 2009, 28(3): 57:1–57:8.

[9] Salimans T, Karpathy A, Chen X, Kingma D P. PixelCNN++: Improving the Pixel-CNN with discretized logistic mixture likelihood and other modifications. In *Proc. the 5th Int. Conf. Learning Representations*, April 2017.

[10] Gregor K, Danihelka I, Graves A, Rezende D, Wierstra D. DRAW: A recurrent neural network for image generation. In *Proc. the 32nd International Conference on Machine Learning*, July 2015, pp.1462–1471.

[11] Kingma D P, Welling M. Auto-encoding variational bayes. In *Proc. the 2nd Int. Conf. Learning Representations*, April 2014.

[12] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In *Proc. the 27th Advances in Neural Information Processing Systems*, December 2014, pp.2672–2680.

[13] Upchurch P, Gardner J, Bala K, Pless R, Snavely N, Weinberger K. Deep feature interpolation for image content changes. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, July 2016, pp.6090–6099.

[14] Perarnau G, Weijer J, Raducanu B, Álvarez J M. Invertible conditional gans for image editing. In *Proc. NIPS Workshop on Adversarial Training*, December 2016.

[15] Yin W, Fu Y, Sigal L, Xue X. Semi-Latent GAN: Learning to generate and modify facial images from attributes. arXiv:1704.02166, 2017. https://arxiv.org/abs/1704.02166, April 2017.

[16] Liao J, Yao Y, Yuan L, Hua G, Kang S B. Visual attribute transfer through deep image analogy. *ACM Trans. Graph.*, 2017, 36(4): 120:1–120:15.

[17] Lu Y, Tai Y W, Tang C K. Conditional CycleGAN for attribute guided face image generation. arXiv: 1705.09966, 2017. https://arxiv.org/abs/1705.09966, May 2017.

[18] Zhu J Y, Park T, Isola P, Efros A A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. International Conference on Computer Vision*, Oct. 2017, pp.2242–2251.

[19] Choi Y, Choi M, Kim M, Ha J W, Kim S, Choo J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. arXiv:1711.09020, 2017. https://arxiv.org/abs/1711.09020, November 2017.

[20] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv: 1411.1784, 2014. https://arxiv.org/abs/1411.1784, Mar. 2018.

[21] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs. In *Proc. the 34th International Conference on Machine Learning*, August 2017, pp.2642–2651.

[22] Isola P, Zhu J Y, Zhou T, Efros A A. Image-to-image translation with conditional adversarial networks. In *Proc. Conference on Computer Vision and Pattern Recognition*, July 2017, pp.5967–5976.

[23] Taigman Y, Polyak A, Wolf L. Unsupervised cross-domain image generation. In *Proc. the 5th Int. Conf. Learning Representations*, April 2017.

[24] Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. In *Proc. International Conference on Computer Vision*, Dec. 2015, pp.3730–3738.

[25] Ronneberger O, Fischer P, Brox T *et al.* U-Net: Convolutional networks for biomedical image segmentation. In *Proc. 18th Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, October 2015, pp.234–241.

[26] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. the 32nd International Conference on Machine Learning*, July 2015, pp.448–456.

[27] Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, June 2013.

[28] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In *Proc. European Conference on Computer Vision*, October 2014.

[29] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado G S, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv: 1603.04467, https://arxiv.org/abs/1603.04467, Mar. 2018.

[30] Kingma D, Ba J. Adam: A method for stochastic optimization. In *Proc. the 3rd Int. Conf. Learning Representations*, May 2015.

[31] Yi Z, Zhang H, Tan P, Gong M. DualGAN: Unsupervised dual learning for image-to-image translation. In *Proc. International Conference on Computer Vision*, Oct. 2017, pp.2868–2876.

[32] Kim T, Cha M, Kim H, Lee J K, Kim J. Learning to discover cross-domain relations with generative adversarial networks. In *Proc. the 34th International Conference on Machine Learning*, August 2017, pp.1857–1865.

Huai-Yu Li is a Ph.D. student in Institute of Automation, Chinese Academy of Sciences, Beijing, under the supervision of Prof. Bao-Gang Hu. He earned his Bachelor's degree in electronic information engineering from Northeast University, Shenyang, in 2014. His research interests are in artificial intelligence, computer vision, and deep learning.

Wei-Ming Dong is a professor in the National Laboratory of Pattern Recognition (NLPR) at Institute of Automation, Chinese Academy of Sciences, Beijing. He received his B.S. and M.S. degrees in computer science in 2001 and 2004, both from Tsinghua University, Beijing. He received his Ph.D. degree in computer science from the University of Lorraine, France, in 2007. His research interests include image synthesis and image recognition. He is a member of CCF, ACM, and IEEE.

Bao-Gang Hu received his M.S. degree from the University of Science and Technology, Beijing, in 1983, and his Ph.D. degree from McMaster University, Canada, in 1993, both in mechanical engineering. From 1994 to 1997, he was a research engineer and senior research engineer at C-CORE, Memorial University of Newfoundland, Canada. Currently, he is a professor with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing. From 2000 to 2005, he was the Chinese director of LIAMA (the Chinese-French Joint Laboratory for Computer Science, Control, and Applied Mathematics). His main research interests include pattern recognition and plant growth modeling. He is a senior member of IEEE and a member of CCF.