# SSCGAN: Facial Attribute Editing via Style Skip Connections

Wenqing Chu[1], Ying Tai[1(✉)], Chengjie Wang[1], Jilin Li[1], Feiyue Huang[1], and Rongrong Ji[2]

[1] Youtu Lab Tencent, Shanghai, China
yingtai@tencent.com
[2] Xiamen University, Xiamen, China

**Abstract.** Existing facial attribute editing methods typically employ an encoder-decoder architecture where the attribute information is expressed as a conditional one-hot vector spatially concatenated with the image or intermediate feature maps. However, such operations only learn the local semantic mapping but ignore global facial statistics. In this work, we focus on solving this issue by editing the channel-wise global information denoted as the style feature. We develop a style skip connection based generative adversarial network, referred to as SSCGAN which enables accurate facial attribute manipulation. Specifically, we inject the target attribute information into multiple style skip connection paths between the encoder and decoder. Each connection extracts the style feature of the latent feature maps in the encoder and then performs a residual learning based mapping function in the global information space guided by the target attributes. In the following, the adjusted style feature will be utilized as the conditional information for instance normalization to transform the corresponding latent feature maps in the decoder. In addition, to avoid the vanishing of spatial details (*e.g.* hairstyle or pupil locations), we further introduce the skip connection based spatial information transfer module. Through the global-wise style and local-wise spatial information manipulation, the proposed method can produce better results in terms of attribute generation accuracy and image quality. Experimental results demonstrate the proposed algorithm performs favorably against the state-of-the-art methods.

**Keywords:** Facial attribute editing · Style feature · Skip connection

## 1 Introduction

Given a facial photo, attribute editing aims to translate the image to enable target attribute transfer while preserving the image content, i.e., the identity information, illumination, and other irrelevant attributes). During the past


A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12360, pp. 414–429, 2020.
https://doi.org/10.1007/978-3-030-58555-6_25

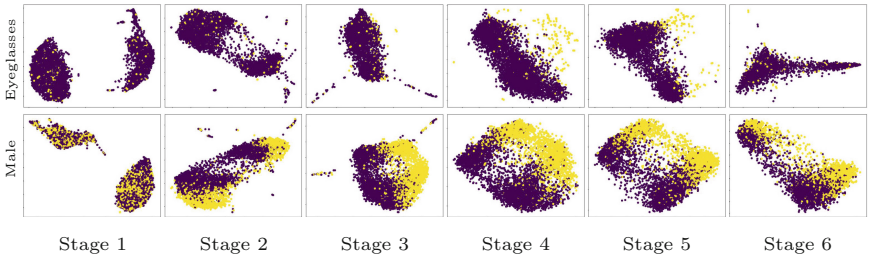


**Fig. 1.** Visualization of style features in different stages for attribute eyeglasses and male. We use t-SNE [27] to visualize the distributions of style features. Each point indicates a sample and different colors indicate whether this sample exhibits this attribute. It is observed that the style features at some specific stages are separated well, which means the style features could represent the attributes accurately. In addition, the style features in stage 4 and stage 3 for eyeglasses and male are the most discriminative, respectively. Therefore, the style features in different stages may capture different attribute information.

decades, attribute guided facial manipulation has drawn considerable attentions [6,12,23,33], and widely used in many real-world visual applications. However, it is very challenging to generate high-quality and accurate facial editing results due to the high-dimensional output space, complex facial structure, and vague attribute definitions. Besides, due to the absence of paired examples for training, it could only be tackled through an unpaired manner resulting in more difficulties.

With the rapid development of Generative Adversarial Networks (GANs) [8], there have been a large number of attempts [5,6,12,23,36] for facial editing. Most existing methods employ an encoder-decoder architecture, and rely on conditional GANs [28,29]. Specifically, the attribute information is represented as a one-hot vector [6], where each bit indicates a specific attribute. This vector is then expanded and concatenated with the input image or intermediate feature maps [6,23,36] to guide the feature transformation. Furthermore, the attribute information will be used as the supervised signals of the auxiliary loss combined with the cycle consistency loss [41] and adversarial loss [8], as to compose the overall objectives for stable and effective training.

However, providing the attributes in each spatial location and then manipulating the feature maps locally may ignore the global structure which leads to unsatisfactory performance. The channel-wise feature map manipulation is an important and effective technique for harvesting the global information in many visual tasks such as image classification [11] and semantic segmentation [40], which has not been well explored in facial attribute editing. That motivated us to perform attribute transfer via the manipulation of the global channel-wise statistics of the latent feature maps.

Following [13,18], we employ the channel-wise mean and variance of the feature maps as the global information and denote them as the *style feature*. Here, we take the advanced image generation method StyleGAN [18] as an example

to verify the relationship between the style feature and different attributes. To be more specific, we leverage an efficient embedding algorithm [1] to compute the style feature of the well-annotated facial attribute dataset CelebA [26] and then employ the Neighborhood Components Analysis [31] to perform supervised dimensionality reduction for each attribute. Then we use t-SNE [27] to visualize the distributions of style features in different stages of the decoder. As shown in Fig. 1, we can observe that the style features at some specific stages are separated well, which means the style features could represent the attributes accurately. In addition, the style features in stage 4 and stage 3 for eyeglasses and male are the most discriminative, respectively. Therefore, the style features in different stages may control different attribute information.

Inspired by the good characteristic of the style feature on controlling facial attributes, we propose to edit the latent feature maps via style skip connections, which modify the global channel-wise statistics to achieve attribute transfer. Specifically, we leverage the style information in the encoder and target attributes to inference the desired statistic information of the latent feature maps in the decoder. Then the manipulated style information is utilized as the conditional input for instance normalization to adjust the distribution of the corresponding latent feature maps in the decoder. However, we find the style information is spatial invariant and may drop the spatial variations, which in some cases describe the local details like the pupil locations or hair texture. To address this issue, we further employ the spatial information based skip connections, which extract the spatial details from the latent feature maps and transfers them to the decoder. In summary, the global-wise style manipulation can handle the facial attribute transfer, and the local-wise spatial information transfer can make up the local finer details.

The main contributions of this work are as follows. First, we introduce a style skip connection based architecture to perform facial attribute editing which manipulates the latent feature maps in terms of global statistic information. Second, a spatial information transfer module is developed to avoid the vanishing of finer facial details. Third, the visual comparisons and quantitative analysis on the large-scale facial attribute benchmark CelebA [26] demonstrate that our framework achieves favorable performance against the state-of-the-art methods.

## 2 Related Work

**Image-to-Image Translation.** Recent years have seen tremendous progress in image-to-image translation, relying on generative adversarial networks [2,8]. To model the mapping from input to output images, Pix2pix [15] utilizes a patch-based adversarial loss which forces the generated images indistinguishable from target images and achieves reasonable results. However, the paired training is usually not available in real-world scenarios, CycleGAN [41], DiscoGAN [19], and UNIT [24] constrain the mapping through an additional cycle consistency loss. Furthermore, MUNIT [14] and DRIT [21] model the input images with disentangled content and attribute representations and thus generate diverse

outputs. In addition, FUNIT [25] handles the few-shot image translation task which only provides a few target images for learning the mapping. However, these methods could only tackle image translation between two domains, thus they can not be applied to the facial attribute transfer task directly.

**Facial Editing.** Most facial editing methods are based on conditional GANs [28,29]. The conditional information can be facial attributes [6,23,36,39], expressions [33,38], poses [3,12] or reference images [5,37]. Among them, facial attribute editing has caused great attentions due to its wide applications. IcGAN [32] generates the attribute-independent latent representation and then the target attribute information is combined as input to the conditional GANs. To achieve better multi-domain translation, StarGAN [6] and AttGAN [9] employ an additional attribute classifier to constrain the output image. Furthermore, STGAN [23] adopts a skip connection based architecture and transfers the feature maps selectively according to the desired attribute change which produces visually realistic editing. Similar to STGAN [23], RelGAN [36] also leverages the relative attribute differences for fine-grained control. Existing methods usually modify the entire feature maps locally according to the desired attributes which ignore the global information. Instead, we find that the statistical information like mean and variance of the feature maps are very informative.

**Style-Based Face Generation.** Recently, a number of improved GANs have been proposed [4,17,18,35] which produce promising results with high resolution. StyleGAN [18] achieves impressive performance by adopting a style-based generator relying on the conditional adaptive instance normalization [13]. That has inspired several extensions [1,34] to perform facial manipulation with StyleGAN. However, these methods [1,34] employ an optimization-based embedding algorithm which uses 5000 gradient descent steps, taking about 7 minutes on an advanced V100 GPU device. Also, they are constrained by the pretrained StyleGAN model and could not be applied to other tasks flexibly.

## 3 Method

In this work, we introduce a facial attribute manipulation algorithm through editing the intermediate feature maps via style and spatial information guided skip connections. As shown in Fig. 1, the style features in multi-stages are responsible for different attributes, respectively. That inspired us to manipulate the facial image by adjusting the global statistic information in the feature maps. Different from existing methods that concatenate the target attribute information with the latent feature maps to achieve local feature transformation, our method aims to edit the facial attributes globally. As a result, the proposed approach can achieve more effective and accurate manipulation.

The overall framework is based on an encoder-decoder architecture shown in Fig. 2. Specifically, we leverage two kinds of skip connections between the

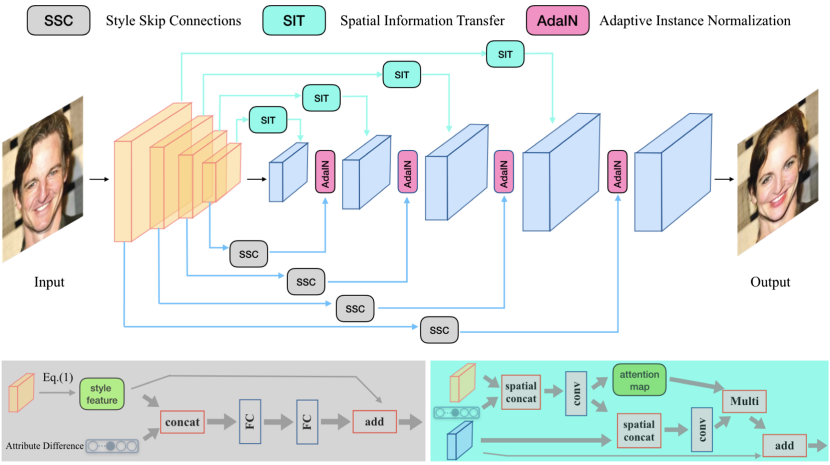


**Fig. 2.** Framework of the proposed algorithm. The overall framework is based on an encoder-decoder architecture. We combine the style skip connections and spatial information transfer to achieve facial attribute editing. Specifically, the style information in the encoder and target attributes are utilized to inference the desired statistic information for adjusting the latent feature maps in the decoder in a global way. Besides, we employ the spatial information based skip connections to transfer spatial details to the decoder, so that the proposed method can achieve local feature transformation and recovery for the pupil locations or hair texture.

encoder and decoder to incorporate the target attribute information. The goal of the first kind of skip connections is to obtain the style information of the latent feature maps in the encoder and then perform a residual learning based style transformation under the instruction of the target attributes. After that, we employ it as the conditional information for instance normalization on the corresponding latent feature maps in the decoder. To make up the vanishing of the facial details, we introduce a spatial information based skip connection which reserves the spatial variations in the latent feature maps. To be more specific, this information will be concatenated with the latent feature maps in the decoder to perform the local feature transformation. In the following, we first describe the network architecture. Next, we describe the style and spatial based image manipulation module, Finally, we present the loss functions for training and implementation details.

### 3.1 Multiple Skip Connections Architecture

Considering the style features in different stages control different attributes as shown in Fig. 1, a skip connection at some specific stage is not general for various attributes. Also, as demonstrated in StyleGAN [18], the style information in the low-resolution feature maps could represent coarse level attributes like pose, face shape and eyeglasses. Instead, the high-resolution feature maps could

control finer details like eyes open/closed, color scheme and micro-structure. As a result, only using part of the skip connections may miss the control for some specific attributes. In addition, the spatial details are also sensitive to different resolutions. Therefore, we utilize multiple skip connections to manipulate both the coarse facial structure and finer local facial information. Specifically, we employ 6 stages for the encoder and decoder, respectively. Each stage in the encoder/decoder has one residual block and down-samples/up-samples the feature maps by 2 times, respectively. Besides, it has the style and spatial based skip connections sent to the corresponding stage in the decoder.

### 3.2 Style Skip Connections

In this section, we introduce the design methodology of the style manipulation part. The goal is to modify the latent feature maps in the decoder globally to enable accurate attribute transfer while preserving irrelevant facial information. Suppose the architectures for the encoder and decoder are symmetric and both have $n$ stages. For simplicity, we denote the feature maps in the encoder and decoder as $\mathbf{f^1_{enc}}$, $\mathbf{f^2_{enc}}$, $\cdots$, $\mathbf{f^n_{enc}}$ and $\mathbf{f^1_{dec}}$, $\mathbf{f^2_{dec}}$, $\cdots$, $\mathbf{f^n_{dec}}$. For $t \in (1, 2, \cdots, n)$, the feature maps $\mathbf{f^t_{enc}}$ and $\mathbf{f^{n+1-t}_{enc}}$ have the same spatial and channel sizes.

We first describe how to represent the channel-wise global information for $f^t_{dec}$. Inspired by the neural style transfer approaches [7,22], we leverage the feature distributions in $f^t_{dec}$ for representing the global facial statistics information. We find the prevalent Gram Matrices of feature maps used in [7] are too large and thus time-consuming compared to the statistics (*i.e.* mean and variance) of Batch Normalization (BN) layers used in [22]. Therefore, we consider using the mean and variance statistics in $\mathbf{f^t_{dec}}$ as the style features for efficiency. Suppose the size of the feature maps $\mathbf{f^t_{dec}}$ is $\mathbb{R}^{N_t \times M_t}$, where $N_t$ is the number of the feature maps in the layer $t$ and $M_t$ is the result of the height multiplying with the width. Similar to [22], we adopt the mean $\mu^t_i$ and standard deviation $\sigma^t_i$ of the $i$-th channel among all the positions of the feature map in the layer $t$ to represent the style:

$$\begin{cases} \mu^t_i = \dfrac{1}{M_t} \displaystyle\sum_{j=1}^{M_t} (\mathbf{f^t_{dec}})_{i,j}. \\ \sigma^{t\,2}_i = \dfrac{1}{M_t} \displaystyle\sum_{j=1}^{M_t} ((\mathbf{f^t_{dec}})_{i,j} - \mu^t_i)^2. \end{cases} \tag{1}$$

Furthermore, we concatenate the $\mu^t$ and $\sigma^t$ into a $(N_t \times 2)$-d vector as the style feature for feature maps $\mathbf{f^t_{dec}}$.

Next, a direct way is to utilize the attribute difference vector as input to generate the style information for adjusting the latent feature maps in the decoder. However, this solution may ignore the original image content and produce incorrect statistic information, which leads to a number of undesired changes in the generated images. To achieve accurate facial attribute editing, we employ the

attribute difference vectors and the style information calculated from the latent feature maps in the encoder stage to produce the desired style information. Note that we find the style features at some specific stage are separated well, which means the style features in different stages could represent different attributes accurately as shown in Fig. 1. Therefore, we perform style information manipulation within the same stage of the encoder and decoder.

In the following, we describe how to perform style information based skip connections between $\mathbf{f_{enc}^{t}}$ and $\mathbf{f_{dec}^{n+1-t}}$. The style feature extracted in $\mathbf{f_{enc}^{t}}$ is concatenated with the attribute difference vector and fed into a 2 layer fully connected neural networks to predict the residual information as shown in Fig. 2. After that, we add it to the original style feature to obtain the desired style information which can be used as the conditional input to manipulate the corresponding $\mathbf{f_{dec}^{n+1-t}}$. Taking efficiency into consideration, we utilize the Adaptive Instance Normalization [13] (AdaIN) to manipulate the global statistic information of the latent feature maps. For all style based skip connections used in the proposed method, we adopt the same embedding way for the desired style feature and network structure.

### 3.3 Spatial Information Transfer

Although the style features could carry most facial information like coarse structure and facial components, the local information may be dropped due to the spatial invariant characteristic of style information and the low resolution of the last encoder stage. For example, the spatial details like the hair texture and pupil locations are very difficult to be embedded into the style features. Therefore, if only use the style feature based skip connections, the generated images may have accurate target attributes but look over smooth. As a result, they are not realistic enough and can not achieve satisfactory performance.

To address the above problem, we develop a spatial information transfer module to collect the spatial details and deliver them to the corresponding latent feature maps. Since the target attribute editing could only need part of the original facial image information, we also provide the attribute difference vector to extract the spatial information more accurately. Specifically, we expand the attribute difference vector spatially and concatenate it with the intermediate feature maps in the encoder, and then we adopt a convolution operation to generate a two-channel feature map. One of them is regarded as representing the spatial details. During the decoder stage, we combine the spatial map with the latent feature maps to predict the residual spatial information. The other one is processed by a sigmoid activation function and then used as an attention map because we want to avoid introducing noise from the residual information through the attention mechanism [33,39]. In the following, the attention map is leveraged to guide the fusion of the original intermediate feature maps and the residual one. Based on the dedicated design, the spatial information transfer module could benefit the editing and recovery of the local spatial details.

### 3.4 Loss Functions

We combine multiple loss functions to train the model in an unpaired manner. To better capture the translation information, we also employ the attribute difference vector $\mathbf{attr}_{diff}$ as the conditional information similar to STGAN [23] and RelGAN [36]. Given an input image $\mathbf{x}$, our framework can generate the output image $\mathbf{y}$ as below:

$$\mathbf{y} = \mathbf{G}(\mathbf{x}, \mathbf{attr}_{diff}). \tag{2}$$

To require the generated image $\mathbf{y}$ satisfying the objective of facial attribute editing, we utilize three constraints: 1) the generated facial image should be the same as input one when the attribute difference is none; 2) the generated facial image should be realistic and similar to the real facial images; 3) the generated image should exhibit the target attributes. Therefore, we employ three loss functions based on the above-mentioned constraints to train the network.

**Reconstruction Loss.** We set the attribute difference vector as $\mathbf{0}$ and fed it with $\mathbf{x}$ into the network to obtain $\mathbf{y_{rec}}$:

$$\mathbf{y_{rec}} = \mathbf{G}(\mathbf{x}, \mathbf{0}). \tag{3}$$

Then we combine the pixel and feature level reconstruction loss as below:

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{x}}\big[\mathcal{L}_1(\mathbf{y_{rec}}, \mathbf{x}) + \mathcal{L}_{perceptual}(\mathbf{y_{rec}}, \mathbf{x})\big], \tag{4}$$

where the perceptual loss $\mathcal{L}_{perceptual}$ introduced in [16] can improve the image quality as demonstrated in [14].

**Adversarial Loss.** In addition, we adopt the adversarial loss [8] which is effective in constraining the generated images looking realistic. The adversarial learning framework consists of two sub-networks, including a generator and a discriminator. Here we leverage the facial attribute editing network as the generator. Given an input image $\mathbf{x}$ and target attribute difference $\mathbf{attr}_{diff}$, our generator can produce the output image $\mathbf{y}$ according to Eq. 2. The discriminator is a fully convolutional neural network and required to distinguish the patches of the real ($\mathbf{x}$) and the generated images ($\mathbf{y}$). Then, the goal of the generator is to fool the discriminator via an adversarial loss denoted as $\mathcal{L}_{adv}$. We employ the same training scheme and loss functions as the Wasserstein GAN model [2] as below:

$$\begin{cases} \mathcal{L}_{\mathbf{dis}} = -\mathbb{E}_{\mathbf{x},\mathbf{attr}_{diff}}\big[\log(1 - \mathbf{D}_{real}(\mathbf{y}))\big] - \mathbb{E}_{\mathbf{x}}\big[\log \mathbf{D}_{real}(\mathbf{x})\big], \\ \mathcal{L}_{\mathbf{adv}} = -\mathbb{E}_{\mathbf{x},\mathbf{attr}_{diff}}\big[\log \mathbf{D}_{real}(\mathbf{y})\big], \end{cases} \tag{5}$$

where minimizing $\mathcal{L}_{dis}$ on the discriminator $\mathbf{D}_{real}$ tries to distinguish between the real and synthesized images. And optimizing $\mathcal{L}_{adv}$ leads to that the generator $\mathbf{G}$ produces visually realistic images.

**Attribute Generation Loss.** To achieve attribute transfer, we utilize an auxiliary attribute generation loss similar to StarGAN [6]. It is achieved by an attribute classifier learned with the real images $\mathbf{x}$ and applied to the generated images $\mathbf{y}$ as a deep image prior. We denote the attribute classification and generation loss functions as below:

$$\begin{cases} \mathcal{L}_{\mathbf{D_{attr}}} = -\sum_{i=1}^{n_{attr}} \left[\mathbf{attr}^i(\log \mathbf{D}^i_{attr}(\mathbf{x})) + (1-\mathbf{attr}^i)\log(1-\mathbf{D}^i_{attr}(\mathbf{x}))\right], \\ \mathcal{L}_{\mathbf{G_{attr}}} = -\sum_{i=1}^{n_{attr}} \left[\mathbf{attr}^i(\log \mathbf{D}^i_{attr}(\mathbf{y})) + (1-\mathbf{attr}^i)\log(1-\mathbf{D}^i_{attr}(\mathbf{y}))\right], \end{cases} \tag{6}$$

where the attribute classifiers $\mathbf{D}_{attr}$ are trained on the real images and optimizing $\mathcal{L}_{G_{attr}}$ aims to require the generated images to satisfy the target attributes.

**Overall Objectives.** The overall objective function for the proposed facial attribute editing network includes the reconstruction/adversarial loss to help generate high quality images, the attribute classification loss to ensure attribute transfer:

$$\mathcal{L}_{overall} = \lambda_r \mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{G_{attr}}. \tag{7}$$

where the hyper-parameters $\lambda_r$ is set to 20.

**Implementation Details.** The proposed framework is implemented with PyTorch [30] and trained with 1 Nvidia V100 GPU. During training, we adopt the Adam [20] optimizer and set the batch size as 32. Similar to CycleGAN [41], we set the initial learning rate as 0.0002 and fix it for the first 100 epochs, and linearly decay the learning rate for another 100 epochs.

## 4 Results and Analysis

In this section, we first describe the basic experiment settings. Next, we perform extensive ablation studies to evaluate different components of the proposed method, including the choices of style and spatial manipulation, embedding manner and multiple skip connections. Finally, we conduct both qualitative and quantitative experiments to compare the proposed algorithm with state-of-the-art methods.

**Datasets.** Following [6,23], we leverage the large scale facial attribution dataset Celeba [26] for evaluation. It contains around 200k facial images and annotates 40 attributes. We randomly select around 180k images for train and validation, and the rest is used as the test set. Besides, we choose 10 attributes to perform facial attribute transfer.

**Table 1.** Attribute generation accuracy for different skip connections.

| Method | Bald | Bangs | Hair | Eyebrow | Glasses | Gender | Mouth | Mustache | Pale | Age | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Spatial | 38.63 | 95.43 | 88.07 | 92.33 | 99.07 | 79.57 | 98.90 | 59.83 | 84.30 | 88.53 | 82.46 |
| Style | 69.60 | **99.93** | 99.83 | **97.97** | **99.97** | 98.20 | 99.87 | 61.83 | 97.13 | 98.30 | 92.26 |
| SSCGAN | **85.40** | 99.23 | **99.30** | 96.57 | 99.93 | **99.10** | **99.90** | **65.73** | **98.03** | **99.00** | **94.21** |

**Evaluation Metrics.** To evaluate the facial attribute editing performance, we take both the attribute generation accuracy and image quality into consideration. Similar to STGAN [23], we utilize the training data to train an attribute classifier and the average attribute classification accuracy on the test set is 95.55%. In all experiments, we use this pretrained classifier to verify the accuracy of facial editing results. In addition, we also follow ELEGANT [37] and RelGAN [36] to employ the Frechet Inception Distance (FID) [10] to demonstrate the image quality. FID aims to evaluate the distribution similarity between two datasets of images. As shown in [10,18], it correlates well to the human evaluation of image quality.

### 4.1 Ablation Study

Here, we investigate the effects of different algorithm designs by comparing the attribute generation accuracy and observing the qualitative results. First, we want to verify the effectiveness of style skip connections.

**Style vs. Spatial.** Based on the encoder-decoder architecture, we adopt the style and spatial skip connections separately to demonstrate their influence. Specifically, we have three settings, including SSCGAN-style, SSCGAN-spatial and SSCGAN (both style and spatial). From Table 1, we can find that SSCGAN-style achieves higher attribute generation accuracy compared with SSCGAN-spatial. Furthermore, employing both kinds of skip connections could obtain the best performance. We also present some qualitative results to demonstrate the editing results. As shown in Fig. 3, we can observe that SSCGAN-spatial does not change the lip color or eyebrow shape as it is not able to learn the global distribution for female appearance. Although SSCGAN-style could change the attributes well, the generated facial images are over smooth and can not maintain some input image information like pupil locations and background. That means the spatial skip connections are also very necessary.

**Embedding Style Information.** Different from the image generation method StyleGAN [18] which utilizes a random noise vector to generate the style information, the facial attribute editing task needs specific style information which combines the input image content and target attributes. Therefore, it is a key challenge to obtain plausible style information. We investigate 5 embedding ways to generate the style information.


**Fig. 3.** An example of editing the attribute gender. We can observe that SSCGAN-spatial does not change the lip color or the eyebrow shape as it is not able to learn the global distribution for female appearance. In addition, SSCGAN-style does not maintain the original spatial information such as the pupil location and background.

- SSCGAN-att: directly leveraging the attribute difference vector to predict the style information through a 2 layer fully connected networks.
- SSCGAN-lm: computing the mean of the last stage in the encoder and concatenating it with attribute difference vector to obtain the style information.
- SSCGAN-lmv: calculating the mean and variance of the last stage and combining it with attribute difference to generate the style information.
- SSCGAN-mm: extracting means for each block in the encoder which are concatenated with attribute difference to predict the style information for the corresponding block in the decoder.
- SSCGAN: utilizing a residual learning based network to generate the style information for each block in the decoder.

**Table 2.** Attribute generation accuracy for different embedding settings.

| Method | Bald | Bangs | Hair | Eyebrow | Glasses | Gender | Mouth | Mustache | Pale | Age | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SSCGAN-att | **88.40** | 98.63 | **99.43** | 87.47 | 99.73 | 85.60 | 99.67 | 51.17 | 95.77 | 94.43 | 90.03 |
| SSCGAN-lm | 52.20 | 98.10 | 96.33 | 93.23 | 99.83 | 89.67 | 99.80 | 52.00 | 95.33 | 94.40 | 87.08 |
| SSCGAN-lmv | 61.40 | 98.57 | 96.60 | 93.33 | 99.53 | 93.07 | 99.60 | 47.90 | 94.60 | 96.17 | 88.07 |
| SSCGAN-mm | 55.73 | 98.63 | 98.77 | 95.37 | 99.63 | 92.17 | 99.57 | 49.13 | 93.40 | 93.63 | 87.60 |
| SSCGAN | 85.40 | **99.23** | 99.30 | **96.57** | **99.93** | **99.10** | **99.90** | **65.73** | **98.03** | **99.00** | **94.21** |

We use the same experiment setting to train these variants and show their performance in Table 2. We observe that the proposed SSCGAN achieves the best attribute generation accuracy. And embedding style information with the feature maps in different stages can surpass only using the last one. In addition, the

**Table 3.** Attribute generation accuracy for different layers.

| Method | Bald | Bangs | Hair | Eyebrow | Glasses | Gender | Mouth | Mustache | Pale | Age | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SSCGAN-8 | 49.50 | **99.47** | 99.43 | 97.27 | 99.83 | 96.93 | 99.53 | 47.73 | 98.37 | 94.83 | 88.28 |
| SSCGAN-16 | 71.57 | 99.47 | **99.93** | **97.93** | 99.97 | 97.77 | 99.90 | 74.27 | **98.80** | 96.77 | 93.63 |
| SSCGAN-32 | 55.50 | 99.50 | 98.60 | 95.40 | 99.90 | 95.10 | 99.87 | **68.70** | 94.43 | 96.20 | 90.32 |
| SSCGAN-64 | 34.60 | 97.93 | 95.33 | 93.83 | 99.83 | 87.37 | 99.47 | 45.57 | 90.43 | 88.70 | 83.30 |
| SSCGAN-128 | 34.17 | 96.07 | 88.63 | 88.37 | 98.33 | 80.70 | 98.17 | 24.67 | 84.80 | 84.00 | 77.79 |
| SSCGAN | **85.40** | 99.23 | 99.30 | 96.57 | **99.93** | **99.10** | **99.90** | 65.73 | 98.03 | **99.00** | **94.21** |

**Table 4.** Comparisons of different methods on the attribute generation accuracy.

| Method | Bald | Bangs | Hair | Eyebrow | Glasses | Gender | Mouth | Mustache | Pale | Age | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| StarGAN | 13.30 | 93.20 | 68.20 | 84.05 | 94.96 | 75.60 | 98.94 | 12.23 | 75.01 | 86.07 | 70.15 |
| AttGAN | 21.20 | 89.80 | 76.27 | 68.17 | 98.17 | 68.03 | 95.43 | 18.87 | 87.07 | 70.03 | 69.30 |
| STGAN | 58.93 | 99.23 | 87.27 | 95.07 | 99.37 | 73.34 | 98.70 | 45.20 | 96.89 | 78.13 | 83.21 |
| RelGAN | 51.39 | 96.50 | 98.33 | 72.33 | 99.10 | 99.60 | 85.57 | 45.37 | 91.97 | 95.83 | 83.59 |
| SSCGAN | **85.40** | v99.23 | **99.30** | **96.57** | **99.93** | **99.10** | **99.90** | **65.73** | **98.03** | **99.00** | **94.21** |

**Table 5.** Comparisons of different methods on the FID scores.

| Method | StarGAN | AttGAN | STGAN | RelGAN | Ours |
|---|---|---|---|---|---|
| FID | 14.27 | 6.82 | 4.78 | 5.13 | **4.69** |

usage of both mean and variance information is helpful as SSCGAN-lmv obtains better results than SSCGAN-lm. In summary, generating style information in a residual learning manner for each style skip connection is the best way.

**Multiple Skip Connections.** Furthermore, we are interested in the influence of multiple skip connections. Specifically, we investigate to only use a single skip connection in the network architecture. Therefore, we can obtain 5 variants which only perform feature manipulation at $8 \times 8$, $16 \times 16$, $32 \times 32$, $64 \times 64$, $128 \times 128$ scale level which are denoted as SSCGAN-8, SSCGAN-16, SSCGAN-32, SSCGAN-64, SSCGAN-128. From Table 3, we can find that only using specific skip connection degrades the overall performance. In addition, the experimental results demonstrate that different scale level manipulations have different effects on the performance of attribute editing.

### 4.2 Comparisons with State-of-the-Arts

In the following, we compare the proposed framework with several state-of-the-art methods. We follow the pioneering STGAN [23] and RelGAN [36] to perform quantitative and qualitative experimental evaluations.

**Baselines.** The recently proposed StarGAN [6], AttGAN [9], STGAN [23] and RelGAN [36] are used as the competing approaches. They all use the encoder-decoder architecture and the overall objectives are also similar. To compare these


**Fig. 4.** An example of editing the attribute bangs. Existing methods all incorporate the attribute information through concatenating it with the feature maps. That may lead to inaccurate changes or appearance inconsistent. Our method based on global style manipulation could achieve better visual results.

existing methods under the same experimental setting including train/validation data split, image cropping manner, image resolution and, selected attributes, we use the official released codes and train these models under their default hyper-parameters. We find that the performance of the state-of-the-art methods AttGAN, STGAN and RelGAN on the attribute generation accuracy is close to those reported in the original paper. Therefore, the following comparisons are fair and convincing.

**Quantitative Results.** From Table 4, we can observe that the proposed method achieves the best average attribute generation accuracy (94.21%). STGAN [23] and RelGAN [36] leverages attribute difference vectors as conditional information, thus their results are better than StarGAN [6] and AttGAN [9]. However, they all introduce the attribute information locally by concatenating it with the intermediate feature maps in each spatial location, which leads to unsatisfactory editing performance. In contrast, our method is able to learn global appearances for different attributes which results in more accurate editing results. Furthermore, we compare the editing performance of these methods in terms of FID scores which can indicate the image quality well. Here, we provide FID scores for all generated images in Table 5. The experimental results demonstrate that our method performs favorable against existing facial attribute editing approaches.

**Qualitative Results.** In addition, we show an example to illustrate the facial editing performance for bangs of different methods in Fig. 4. The proposed style skip connections aim to manipulate the feature maps in a global channel-wise manner, and thus both input and output of the style skip connections are *compact vectors* which represent high-level semantics. In contrast, spatial concatenation learns the mapping on complex local regions which is a more difficult scenario

than on the channel-wise vectors. As shown in the first row in Fig. 4, StarGAN modifies the irrelevant facial region and RelGAN produces inconsistent bangs compared with the hair. For the second row in Fig. 4, the results of AttGAN, STGAN and RelGAN are not correct around the hair. Furthermore, we show an example of the facial editing results for multiple attributes in Fig. 5.

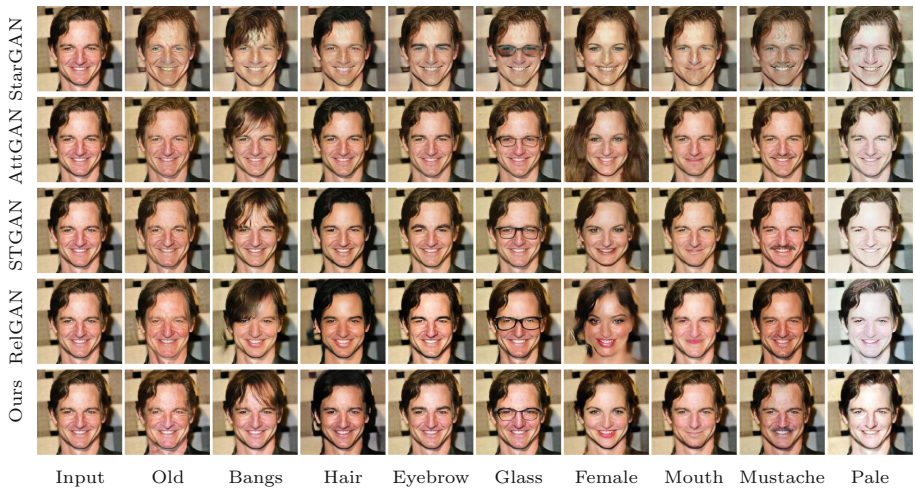


**Fig. 5.** Results of different facial attribute editing methods. Existing methods all introduce the attribute information locally, which leads to unsatisfactory editing performance. Instead, through the global-wise style and local-wise spatial information manipulation, the proposed method can achieve favorable performance for most attributes.

## 5 Conclusions

In this work, we introduce a style skip connection based encoder-decoder architecture for facial attribute editing. To incorporate the target attributes with the image content, we propose to edit the statistics information of the intermediate feature maps in the decoder according to the attribute difference. The manipulation in the style space could translate the facial image in a global way which is more accurate and effective. Furthermore, a spatial information transfer module is developed to avoid the vanishing of the spatial details. In experiments, visual comparisons and quantitative results demonstrate that our method can generate accurate and high-quality facial results against state-of-the-art methods. In the future, we will investigate to apply the proposed algorithm to other visual tasks such as semantic segmentation, image colorization, to name a few.

## References


1. Abdal, R., Qin, Y., Wonka, P.: Image2StyleGAN: how to embed images into the StyleGAN latent space? In: ICCV (2019)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: ICML (2017)
3. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Towards open-set identity preserving face synthesis. In: CVPR (2018)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2019)
5. Chang, H., Lu, J., Yu, F., Finkelstein, A.: Pairedcyclegan: asymmetric style transfer for applying and removing makeup. In: CVPR (2018)
6. 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. In: CVPR (2018)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR (2016)
8. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS (2014)
9. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: AttGAN: facial attribute editing by only changing what you want. IEEE Trans. Image Process. **28**(11), 5464–5478 (2019)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS, pp. 6626–6637 (2017)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
12. Huang, R., Zhang, S., Li, T., He, R.: Beyond face rotation: global and local perception gan for photorealistic and identity preserving frontal view synthesis. In: ICCV, pp. 2439–2448 (2017)
13. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)
14. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
17. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: ICLR (2018)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
19. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: ICML (2017)
20. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
21. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M.K., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: ECCV (2018)
22. Li, Y., Wang, N., Liu, J., Hou, X.: Demystifying neural style transfer (2017)
23. Liu, M., et al.: STGAN: a unified selective transfer network for arbitrary image attribute editing. In: CVPR (2019)

24. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NIPS (2017)
25. Liu, M.Y., et al.: Few-shot unsupervised image-to-image translation. In: ICCV (2019)
26. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
27. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(Nov), 2579–2605 (2008)
28. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
29. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: ICML (2017)
30. Paszke, A., et al.: Automatic differentiation in PyTorch (2017)
31. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
32. Perarnau, G., Van De Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional GANs for image editing. arXiv preprint arXiv:1611.06355 (2016)
33. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: anatomically-aware facial animation from a single image. In: ECCV (2018)
34. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of GANs for semantic face editing. arXiv preprint arXiv:1907.10786 (2019)
35. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: CVPR (2018)
36. Wu, P.W., Lin, Y.J., Chang, C.H., Chang, E.Y., Liao, S.W.: ReLGAN: multi-domain image-to-image translation via relative attributes. In: ICCV (2019)
37. Xiao, T., Hong, J., Ma, J.: ELEGANT: exchanging latent encodings with GAN for transferring multiple face attributes. In: ECCV (2018)
38. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. arXiv preprint arXiv:1905.08233 (2019)
39. Zhang, G., Kan, M., Shan, S., Chen, X.: Generative adversarial network with spatial attention for face attribute editing. In: ECCV (2018)
40. Zhang, H., et al.: Context encoding for semantic segmentation. In: CVPR (2018)
41. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)