Spatially-invariant Style-codes Controlled Makeup Transfer

Han Deng[1], Chu Han[2,*], Hongmin cai[1], Guoqiang Han[1], Shengfeng He[1†]
[1] School of Computer Science and Engineering, South China University of Technology
[2] Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences

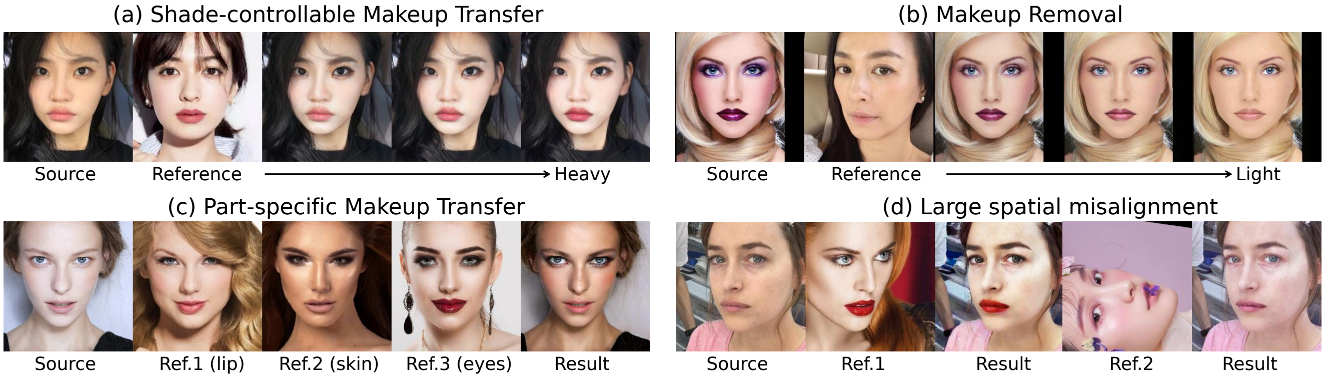


Figure 1. We propose a style-codes controlled makeup transfer method that allows flexible makeup editing without the need for well-aligned source-reference pair. (a) We control the shade of makeup by linear interpolation of two style-codes from the source and the reference images. (b) Makeup removal can be achieved by simply swapping the source and the reference images. (c) We produce partial-transferred results by integrating the style-codes of three reference images. (d) Robust results can be generated with large spatial misalignment.

Abstract

Transferring makeup from the misaligned reference image is challenging. Previous methods overcome this barrier by computing pixel-wise correspondences between two images, which is inaccurate and computational-expensive. In this paper, we take a different perspective to break down the makeup transfer problem into a two-step extraction-assignment process. To this end, we propose a Style-based Controllable GAN model that consists of three components, each of which corresponds to target style-code encoding, face identity features extraction, and makeup fusion, respectively. In particular, a Part-specific Style Encoder encodes the component-wise makeup style of the reference image into a style-code in an intermediate latent space W. The style-code discards spatial information and therefore is invariant to spatial misalignment. On the other hand, the style-code embeds component-wise information, enabling flexible partial makeup editing from multiple references. This style-code, together with source identity features, is integrated into a Makeup Fusion Decoder equipped with multiple AdaIN layers to generate the final result. Our proposed method demonstrates great flexibility on makeup transfer by supporting makeup removal, shade-controllable makeup transfer, and part-specific makeup transfer, even with large spatial misalignment. Extensive experiments demonstrate the superiority of our approach over state-of-the-art methods. Code is available at https://github.com/makeuptransfer/SCGAN.

1. Introduction

Makeup is one of the best ways to make people attractive. However, applying makeup is time-consuming and it even takes a much longer time to seek suitable makeup for each individual. Thus, it is practical to automatically transfer the makeup from reference images onto our own faces. Deep learning approaches, especially GAN-based models, have been widely exploited in makeup transfer task [19, 1, 6, 2]. They mainly model the makeup transfer and recovery processes as a closed-loop to combat the problem of lacking paired data.

Notwithstanding the demonstrated success, these methods are constrained by the input condition, *i.e.*, both the source and reference images must be well-aligned frontal faces. This is because enforcing cycle-consistency [29] cannot guarantee correct spatial transformation. On the other hand, their CycleGAN-based solutions [1, 19] lack flexibility and controllability of makeup styles. A recent work, PSGAN [13], is proposed to address these problems. It computes the dense correspondence attention between two images to consolidate component-to-component transfer. However, apart from the large computational overhead of pixel-wise correspondence, PSGAN suffers from two main issues. First, the predicted pixel-to-region attention is

[*] The first two authors contributed equally.
[†] Corresponding author (hesfe@scut.edu.cn).

ambiguous and thus leading to the color bleeding problem around facial components (see results in Sec. 4). Second, it is cumbersome to implement local transfer from multiple references, as it requires computing dense correspondences for every image and reconstructs a new attentive matrix in a pixel-wise manner.

In this paper, we overcome the spatial misalignment barrier from a completely different perspective. We aim to extract the spatially invariant 1D style-code from the reference image and re-assign it to the source one. Our two-step principle gets over the challenging pixel-wise matching, leading to a simple learning emphasis of makeup assignment. To this end, we propose a new model, called Style-based controllable GAN (SCGAN), that consists of two "extraction" and one "assignment" modules. Particularly, a Part-specific Style Encoder (PSEnc) is designed to extract the makeup style of each part (*e.g.*, lip, skin, and eyes) from one (or multiple) face(s) with the given face parsing maps. The extracted style-code is encoded in a component-wise manner, enabling flexible local manipulation in the downstream applications. On the other hand, inspired by StyleGAN [15], the makeup style is mapped into an intermediate style space instead of using the linearly projected vectors, thus the extracted attributes are less entangled to different factors of variation. Meanwhile, we propose a Face Identity Encoder (FIEnc) to extract the face identity features from the source image. Then, a Makeup Fusion Decoder (MFDec) is presented to progressively fuses the style-code and the face identity features in different feature levels using AdaIN layers [11] and generates the final makeup transfer results. Our proposed model demonstrates great flexibility in makeup transfer, as shown in Fig. 1. Our model achieves shade-controllable makeup transfer by linear combining the style-codes from the source and the reference images (Fig. 1(a) and (b)). Partial transfer from different reference images can be easily achieved by integrating their style-codes (Fig. 1(c)). More importantly, our proposed model is invariant to pose variations (Fig. 1(d)). Extensive experiments and comparisons demonstrate the superiority of our proposed model comparing with the state-of-art approaches.

Our contributions are three-fold:

- We propose a fully automatic makeup transfer model with the best flexibility comparing with existing approaches. Global/local makeup transfer and removal with shade-control can be easily realized by editing the style-codes without extra computational efforts.

- We break down the makeup transfer problem into a two-step extraction-assignment process. A style-based network PSEnc is proposed to map the makeup style into a component-wise style-code. This design eliminates the spatial misalignment problem.

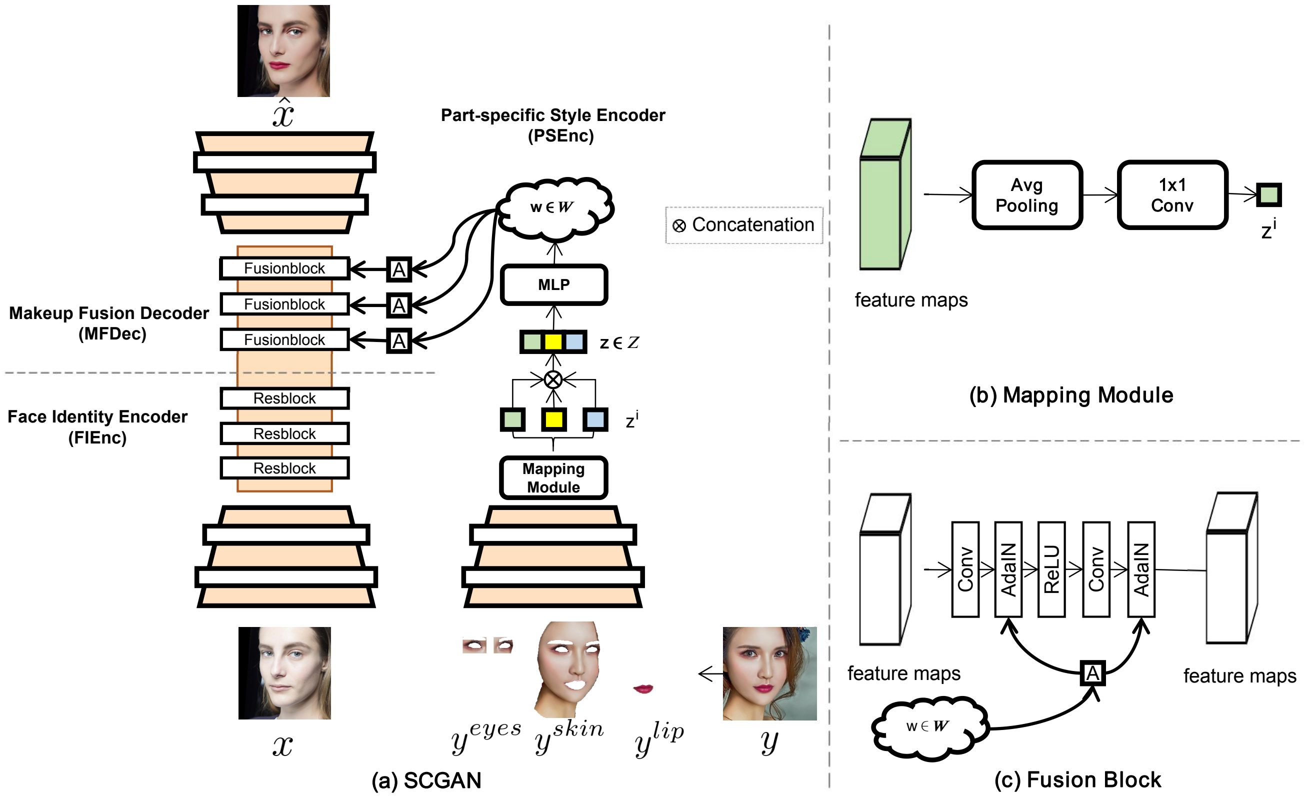
- Our proposed model achieves state-of-the-art performance even with large spatial misalignment between the source and the reference images.

2. Related Work

Makeup transfer aims to transfer a specific makeup style of one face to another. It has been studied for a decade [7, 26, 21, 24, 18]. CycleGAN [29] is one of the most inspiring approaches for makeup transfer, as it is designed to perform image-to-image translation between two unpaired images. However, it can not specify a reference image. PairedCycleGAN [1] further introduces an asymmetric function to complete the task of makeup transfer/removal and variants of cycle consistency losses to support makeup transfer using a specific makeup image. BeautyGlow [2] leverages Glow framework [17] to disentangle the latent features into makeup features and non-makeup features, and then invert the recombined features to image domain. BeautyGAN [19] introduces a dual input/output GAN to complete makeup transfer and makeup removal simultaneously and a makeup loss to refine local details. Local Adversarial Disentangling Network [6] utilizes multiple overlapping local discriminators and asymmetric loss functions to ensure local details consistency. Although the above methods can perform makeup transfer in some senses, they do not specifically handle the spatial misalignment problem between the source image and the reference image.

PSGAN [13] proposes the first attempt to explicitly solve the spatial misalignment problem by introducing an attention mechanism [25]. They build the pixel-wise correspondences and achieve partial makeup transfer by leveraging the face parsing masks and facial landmarks. However, as discussed in Sec. 1, their method relies on the ambiguous pixel-to-region attention. On the other hand, the computed attentive matrix is computational-expensive while not flexible for local transfer. On the contrary, the proposed style-code design enables a large degree of flexibility and controllability. More importantly, our proposed method is invariant to spatial variations of faces.

Style transfer can be regarded as a general form of makeup transfer, and it has been a popular topic in recent years. People tried to transfer a source image to all kinds of styles, such as Van Gogh, pixel arts, cartoon style, etc. Traditional style transfer can be classified into following three categories: (1) Stroke-based rendering [9]; (2) Region-based techniques [23, 4]; (3) Example-based rendering [10, 28]. With the development of the neural network, CNN-based models have been exploded rapidly [3, 11, 14]. However, these style transfer approaches generally transfer the style from one domain to another. They lack local-understanding and controllability, and thus cannot suit face-specific makeup transfer applications.



(a) SCGAN

(b) Mapping Module

(c) Fusion Block

Figure 2. The overview of the proposed method (SCGAN). In (a), the reference image y is decomposed into three parts. Part-specific style encoder extracts the features of each part and maps them into a disentangled style latent space W. Face identity encoder extracts the face identity features of the source image x. Makeup Fusion Decoder fuses the style-code w with the face identity feature to generate final result $\hat{x}$. (b) shows the mapping module of PSEnc. (c) is the fusion block equipped with the AdaIN layer in MFDec.

3. Approach

3.1. Formulation

Let $X = \{x_n | x_n \in X\}_{n=1,\ldots,N}$ and $Y = \{y_m | y_m \in Y\}_{m=1,\ldots,M}$ denote the non-makeup image domain and the makeup image domain, respectively. Given a source image x and a reference image y, we aim to learn the mapping function between two domains: $\hat{x} = \mathcal{G}(x, y)$ and $\hat{y} = \mathcal{G}(y, x)$. $\hat{x}$ is the transferred result with the makeup style of y and the face identity of x.

Fig. 2 depicts the systematic design of our proposed SCGAN. It consists of three network components, a Part-specific Style Encoder (PSEnc), a Face Identity Encoder (FIEnc), and a Makeup Fusion Decoder (MFDec). PSEnc extracts the features of the reference image and maps them into a 1D style-code w. We decompose the reference image into three major components and feed them into PSEnc one-by-one. FIEnc extracts the face identity features of the source image. MFDec fuses the makeup style and the face identity features and then generates a makeup transferred result. The details of the network structure are demonstrated in Sec. 3.2. The objective functions of the complete model can be found in Sec. 3.3.

3.2. Network Structure

3.2.1 Part-specific Style Encoder

As a crucial part of our two-step pipeline, we first delve into the extraction of reference makeup style. Although style-codes can be obtained by simply averaging facial features into a 1D vector (see Fig. 2(b)), the obtained codes follow a similar probability density of training data, leading to unavoidable entanglement between facial components. We present two key strategies to handle this issue. First, inspired by StyleGAN [15], we introduce a non-linear mapping network to embed the initial style-code into an intermediate style latent space. By doing this, the generated style-code is not restricted from training data distribution and is therefore allowed to be disentangled. Second, to introduce prior disentangled knowledge and further support part-specific makeup transfer with greater controllability, we decompose each reference face into three parts (lip, skin, and eyes) by applying a face parser [27]

$$y^i = y \odot M^i. \tag{1}$$

We denote each component of the reference face as y^i, where $i = \{lip, skin, eyes\}$, M^i is the corresponding mask and $\odot$ denotes the Hadamard product. Each input component y^i is fed into a feature extractor with two downsampling convolutional layers. After a mapping module (Fig. 2 (b)) with an average pooling layer and a 1×1 convolutional layer, we map each component y^i to a part-specific style-code z^i. We concatenate the codes of three components and form a complete initial style-code z in Z latent space

$$z = z^{lip} \otimes z^{skin} \otimes z^{eyes}, \tag{2}$$

where $\otimes$ denotes the concatenation. Since we decompose the input reference image into several semantic parts, our proposed SCGAN supports composing arbitrary semantics part from different reference images even with different poses and face expressions

$$z = z_a^{lip} \otimes z_b^{skin} \otimes z_c^{eyes}, \tag{3}$$

where a, b, c indicate three reference images y_a, y_b, y_c respectively.

To get rid of the training data distribution and inject non-linearity, we feed the initial style-code z into a multi-layer perceptron (MLP) with three fully connected layers to map the style-code z into a code w in W latent space with better feature disentanglement

$$w = \text{MLP}(z). \tag{4}$$

3.2.2 Face Identity Encoder

FIEnc serves for face identity feature extraction. It consists of two downsampling convolutional layers and three resblocks. FIEnc takes the input source image x and extract the face identity features $\mathcal{F}_{id}$

$$\mathcal{F}_{id} = \text{FIEnc}(x). \tag{5}$$

Note that, three resblocks in FIEnc are all common residual blocks [8] without AdaIN layers.

3.2.3 Makeup Fusion Decoder

MFDec fuses the style-code w with the face identity features $\mathcal{F}_{id}$ progressively and applies the makeup style from reference image to the face of the source image. To be specific, it consists of three fusion blocks and two upsampling convolutional layers. We introduce two AdaIN layers for each fusion block (Fig. 2 (c)) in MFDec. The style-code w is specialized by a learnable affine transform and then passed into each fusion block. The j-th AdaIN layer is defined as follows:

$$\text{AdaIN}(\mathcal{F}_j, w_j) = w_{s,j} \frac{\mathcal{F}_j - \mu(\mathcal{F}_j)}{\sigma(\mathcal{F}_j)} + w_{b,j}, \tag{6}$$

where $w_{s,j}$ and $w_{b,j}$ are the scaled and biased style using corresponding scalar components, $\mathcal{F}_j$ denotes input feature maps, $\mu(\cdot)$ and $\sigma(\cdot)$ are the channel-wise mean and standard deviation respectively. After two upsampling convolutional layers, we can finally obtain the result $\hat{x}$.

3.3. Full Objective

Since the makeup images and the non-makeup images are unpaired, we trained the network in a cyclic manner. Given the non-makeup domain X and the makeup domain Y, our proposed SCGAN has to learn the mapping bidirectionally ($X \to Y$ and $Y \to X$) between these two domains. So the complete training process is like a CycleGAN [29].

Adversarial loss. Adversarial loss is introduced to guide SCGAN (generator) for more realistic results. We applied two discriminator $\mathcal{D}_X$, $\mathcal{D}_Y$ to discriminate fake or real of the images in X and Y, respectively. The structures of two discriminators are the same as the Markovian discriminator proposed by [12]. Adversarial loss [5] $\mathcal{L}_{\mathcal{G}}^{GAN}$ for generator and $\mathcal{L}_{\mathcal{D}}^{GAN}$ for discriminators are defined as follows

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}^{GAN} = &-\mathbb{E}_{x\sim X}[log\mathcal{D}_X(x)] - \mathbb{E}_{y\sim Y}[log\mathcal{D}_Y(y)] \\ &- \mathbb{E}_{x\sim X, y\sim Y}[log((1 - \mathcal{D}_X(\mathcal{G}(y,x))) \\ &\times (1 - \mathcal{D}_Y(\mathcal{G}(x,y))))],\end{aligned} \tag{7}$$

$$\mathcal{L}_{\mathcal{G}}^{GAN} = -\mathbb{E}_{x\sim X, y\sim Y}[log(\mathcal{D}_X(\mathcal{G}(y,x)) \times \mathcal{D}_Y(\mathcal{G}(x,y)))]. \tag{8}$$

Global perceptual loss. Since the images are from the two domains, pixel-level constraint is not available. To guarantee the face identity between the input source image and the output transferred image, we use a perceptual loss [14] to maintain the global face identity. $\mathcal{L}_{global}^{vgg}$ is defined as

$$\begin{aligned}\mathcal{L}_{global}^{vgg} = &\left\|F_l(\mathcal{G}(y,x)) - F_l(y)\right\|_2 \\ &+ \left\|F_l(\mathcal{G}(x,y)) - F_l(x)\right\|_2,\end{aligned} \tag{9}$$

where $F_l(\cdot)$ denotes the features of l-th layer on the VGG [22] model and $\|\cdot\|_2$ is the L2-Norm.

Local perceptual loss. Besides global perceptual loss, local perceptual loss is introduced to further keep the non-transferred parts unchanged, *e.g.*, teeth, eyebrows, etc. $\mathcal{L}_{local}^{vgg}$ is defined as

$$\begin{aligned}L_{local}^{vgg} = &\sum_{i=1}^{I} \left\|F_l(\mathcal{G}(y,x) \odot M_{y,i}) - F_l(y \odot M_{y,i})\right\|_2 \\ &+ \sum_{i=1}^{I} \left\|F_l(\mathcal{G}(x,y) \odot M_{x,i}) - F_l(x \odot M_{x,i})\right\|_2,\end{aligned} \tag{10}$$

where M denotes the mask of the specific part in $I = \{teeth, hair, eyeballs, eyebrows\}$, i denotes the index of I.

Cycle consistency loss. For unsupervised learning with unpaired images, we use the cycle consistency loss [29], which is defined as

$$\begin{aligned}\mathcal{L}_{cyc} = &\left\|\mathcal{G}(\mathcal{G}(y,x),y) - y)\right\|_1 \\ &+ \left\|\mathcal{G}(\mathcal{G}(x,y),x) - x)\right\|_1,\end{aligned} \tag{11}$$

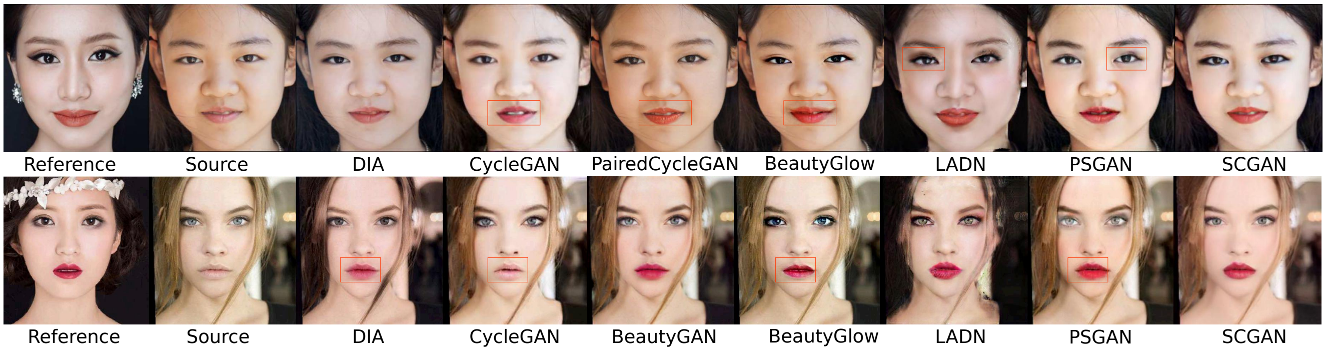


Figure 3. Qualitative comparisons with existing models without spatial misalignment. The highlighted areas are some unnatural or incorrect transfer results.

where $\|\cdot\|_1$ denotes the L1-Norm.

Makeup loss. Makeup loss was proposed by Li *et al.* [19] which utilizes Histogram Matching (HM) to provide a transferred image as a pseudo ground truth. It consists of local histogram matching on 3 different facial regions: skin, lip, and eyes. Then the three transferred parts are integrated as a pseudo ground truth. The makeup loss $\mathcal{L}_{makeup}$ is defined as

$$\begin{aligned}\mathcal{L}_{makeup} = &\left\|\mathcal{G}(x,y) - HM(x,y)\right\|_2 \\ &+ \left\|\mathcal{G}(y,x) - HM(y,x)\right\|_2,\end{aligned} \tag{12}$$

where $HM(\cdot)$ denotes the histogram matching and the output of $HM(x,y)$ has the makeup style of y while preserving the identity of x.

Total loss. The total loss $\mathcal{L}_{total}$ of the complete network is defined as

$$\begin{aligned}\mathcal{L}_{total} = &\lambda_{GAN}(\mathcal{L}_{\mathcal{D}}^{GAN} + \mathcal{L}_{\mathcal{G}}^{GAN}) + \lambda_{cyc}\mathcal{L}_{cyc} \\ &+ \lambda_g\mathcal{L}_{global}^{vgg} + \lambda_l\mathcal{L}_{local}^{vgg} + \lambda_{makeup}\mathcal{L}_{makeup},\end{aligned} \tag{13}$$

where $\lambda_{GAN}, \lambda_{cyc}, \lambda_g, \lambda_l$, and λ_{makeup} are the weights of the loss terms respectively.

3.4. Implementation Details

Our SCGAN was trained and tested on Makeup Transfer (MT) dataset [19] which contains 3834 female images. There are 1115 non-makeup images and 2719 makeup images including variations in poses, races, and etc. We use the same image split with Li *et al.* [19] that randomly selected 100 non-makeup images and 250 makeup images for testing. The rest of the images are used for training. In all the experiments, images are resized to 256×256. We extract features from $Relu\_4\_1$ layer of VGG16 [22] calculating global perceptual loss and local perceptual loss. The optimizer of the generator and two discriminators is Adam [16] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and a fixed learning rate of 0.0002. The batch size is set to 1.

Method	Property			
	Shade	Part	Misalign.	Landmarks free
BGAN [19]				✓
PGAN [1]				
BGlow [2]	✓			✓
LADN [6]	✓			
PSGAN [13]	✓	✓	✓	
SCGAN	✓	✓	✓	✓

Table 1. Properties of state-of-the-arts makeup transfer methods. "Shade": shade-controllable makeup transfer. "Part": partial transfer. "Misalign.": robust transfer with even large spatial misalignment between the source and the reference images. "Landmarks free": model does not need facial landmarks.

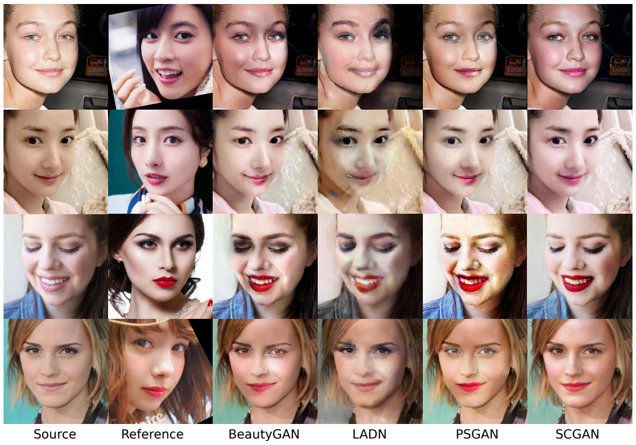


Figure 4. Qualitative comparisons with existing models with large spatial misalignment.

4. Experiments

We compare our SCGAN with two general domain transfer methods: CycleGAN [29] and DIA [20] as well as five state-of-the-art makeup transfer methods: PairedCycleGAN [1], BeautyGAN [19], BeautyGlow [2], LADN [6] and PSGAN [13]. More results can be found in the supplementary materials.

Table 1 summarizes the properties of different makeup transfer methods. Our SCGAN demonstrates the greatest flexibility and controllability among all the models. It is

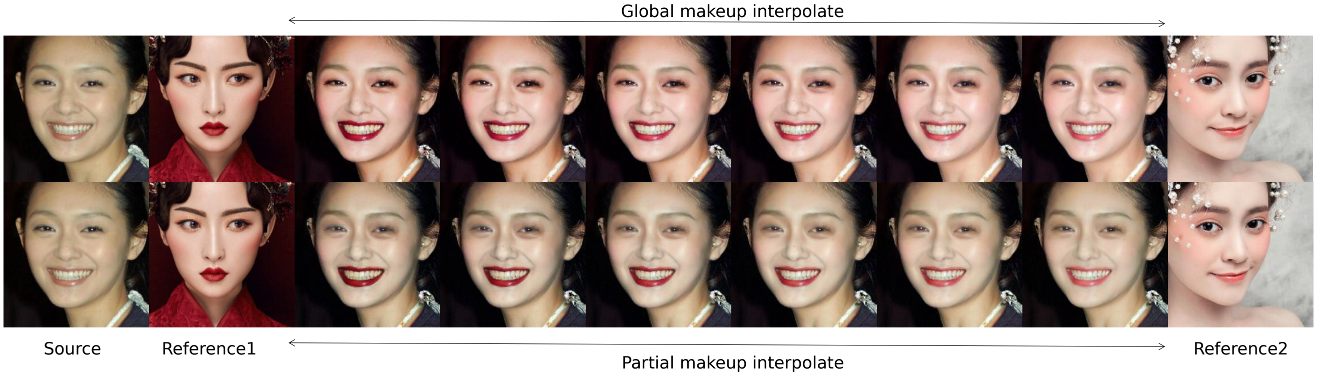


Figure 5. Shade control with two reference images by linear interpolation. The upper row applies a global transfer. The bottom row applies a partial transfer which only transfer the lips style of two reference images.

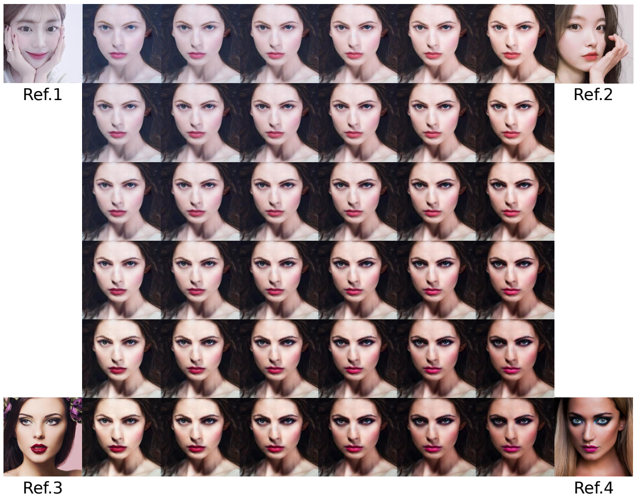


Figure 6. Makeup transfer with four reference images by linear interpolating four style-codes.

a landmark-free model which supports shade-controllable transfer, part-specific transfer, robust transfer even with spatial misalignment.

4.1. Qualitative Comparisons

In Fig. 3, we first show a qualitative comparison with the baseline methods when there is no obvious spatial misalignment between the reference and the source images. We directly copied the results of BeautyGlow and PairedCycleGAN from their papers because they have not released the code. For two domain transfer models, DIA fails to transfer a correct color of the lip and the eyebrows. CycleGAN simply performs domain transfer without any specific style control from the reference image. It synthesizes generally natural results, but it tends to be blurry and lost texture information. On the other hand, methods tailored for makeup transfer also have some weaknesses. There are some severe artifacts in the results of LADN. BeautyGAN and BeautyGlow perform well when there is no spatial misalignment problem. PSGAN can also generate visually acceptable results but locally suffers from color bleeding problem (see highlighted regions), which affects the aesthetic feeling of the results. Meanwhile, PSGAN cannot guarantee a precise transfer while maintaining the source features, *e.g.*, it fails to transfer the correct lip color and lose the source reflection of the lip. In our model, style-codes design solves the spatial misalignment problem and makes sure our model can generate high quality results without the necessity of facial landmark. Local perceptual loss guarantees local correctness and preserves the local textures. Therefore, SCGAN synthesizes the most natural and semantically correct makeup transfer results comparing with existing methods.

To test the robustness against spatial misalignment, we compare our method with three makeup transfer models which have the released code in Fig. 4. When spatial misalignment occurs, LADN and BeautyGAN fail to tackle such a challenge and are not able to provide visually appropriate results. PSGAN produces fewer artifacts than the previous two methods in some cases. However, unnatural shadows are attached to the faces due to the ambiguous attentive matrix. Comparing with the best competitor PSGAN, our SCGAN can generate the most natural results thanks to the spatially invariant style-codes design, leading to a much simpler makeup assignment task. Moreover, we produce pixel-level accurate transfer for each specific part between the source and the reference images, comparing with all the competitors.

4.2. Controllable Makeup Transfer

In this section, we demonstrate the flexibility and controllability of our method, including shade-controllable transfer, part-specific transfer, and makeup removal.

4.2.1 Shade-controllable Transfer

Since we use a style-code w to represent the makeup style of the image. It can easily control the shade by manipulating the style-code. We extract the style-codes from the reference image and the source image. Then we apply a linear interpolation to control the shade (coefficient $\alpha \in [0, 1]$).

$$w = (1 - \alpha)w_{ref} + \alpha w_{source}. \tag{14}$$

Fig. 1 (a) demonstrates the results from light to heavy. We can find a gradual change of makeup style from the source image to the reference image.

Meanwhile, SCGAN also supports fusing the styles from multiple reference images with linear interpolation.

$$w = (1 - \alpha)w_{ref1} + \alpha w_{ref2}. \tag{15}$$

The upper row of Fig. 5 shows the results with different contributions of two reference images.

Fig. 6 demonstrates a more extreme case with four reference images. Results are shown in a "confusion matrix" style. Even we fuse four style-codes from different reference images, the results are natural and the image quality of the results is consistently high. If one style-code contributes more, the makeup style of the result is closer to the reference image of this style-code.

4.2.2 Part-specific Transfer

Since we decompose each reference image into three major components (lip, eyes and skin), SCGAN can support part-specific makeup transfer by simply choosing any specific part from any reference image we want. Given a source image x and three reference images y_1, y_2, and y_3, if we want the lip, eyes, and skin are from y_1, y_2, and y_3 respectively, we can achieve that by passing each specific part to PSEnc as follows

$$z = z_1^{lip} \otimes z_2^{skin} \otimes z_3^{eyes}. \tag{16}$$

Fig. 1 (c) demonstrates the result of the above scenario. We can find that the result is now with the lip from "ref1", skin from "ref2", and eyes from "ref3". The bottom row of Fig. 5 also shows a partial transfer on the lips of the reference images. So w_{ref1} and w_{ref2} are extracted from $\{y_1^{lip}, x^{skin}, x^{eye}\}$ and $\{y_2^{lip}, x^{skin}, x^{eye}\}$.

4.2.3 Makeup Removal

Since our SCGAN learns a bi-directional mapping between two domains in a cyclic manner, we can achieve makeup removal by interchanging the roles of $x \in X$ and $y \in Y$. That is to let the non-makeup image x be the reference image and the makeup image y be the source image. Fig. 1(b) shows makeup removal results from heavy to light.

4.3. Spatial Invariance

Thanks to the style-code design, the transfer results are invariant to spatial and rotational information. This is a practical property for makeup transfer as it frees the users from providing frontal sources/references, and it is an essential factor for applying it in live streaming. To prove this property, we manually introduce rotations to the source image and the reference image and then compare our SCGAN with the state-of-the-art method PSGAN in Fig. 7. PSGAN performs well when the source image and the reference image are with the same rotation. However, when the rotations between them are different, color bleeding problem occurs in the results of PSGAN because it fails to find the corresponding parts of the source and the reference images. Our SCGAN consistently performs visually appropriate results even with different rotations. We further show the spatial invariance property in video makeup transfer in the supplementary materials.

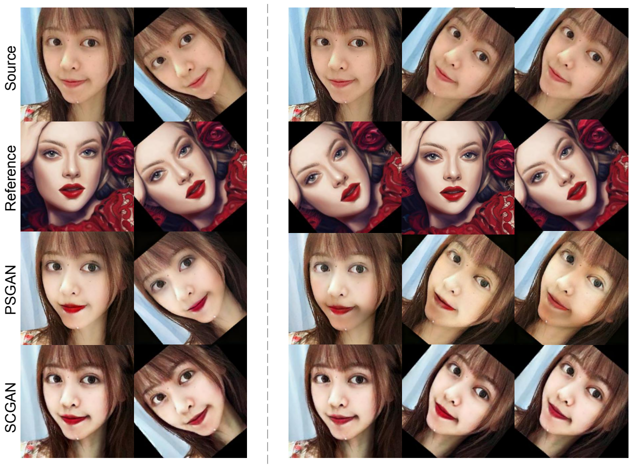


Figure 7. Results with different rotations of the source and the reference images.

4.4. User Study

We conducted three studies to quantitatively evaluate the robustness and the visual quality of our SCGAN comparing with three makeup transfer methods, BeautyGAN, LADN, and PSGAN. In total 30 participants joined these user studies. In the first one, we randomly selected 15 pairs of makeup and non-makeup images which are well-aligned. In the second study, 15 pairs of misaligned images were selected. In the last one, we aim to examine the extreme scenarios of misalignment by selecting 15 pairs of largely rotated photos. In all user studies, participants were asked to select the result with the best visual quality and the most precise transfer. Fig. 8 demonstrates the user studies results, and our SCGAN outperforms all state-of-the-art methods. We believe it is because of our simplified extraction-assignment setting, which leads to accurate transfers in arbitrary scenarios.

4.5. Ablation Study

Local perceptual loss. Besides conventional losses in the existing methods, we introduce an additional local per-

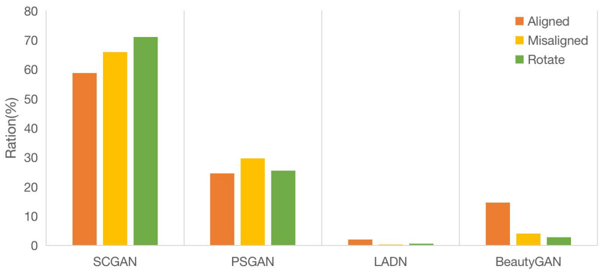


Figure 8. User study results (%). “Aligned” and “Misaligned” indicate the input images are well-aligned or with different poses. “Rotate” means that we randomly rotate the source image or the reference image in “Misaligned” case at a degree ranging from $-45^{\circ} \sim 45^{\circ}$.

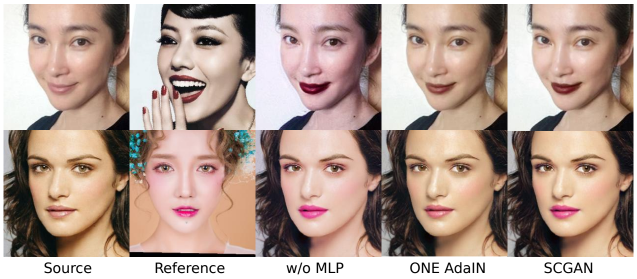


Figure 9. Ablation study of network design. “One AdaIN“ denotes we only adjust the makeup style of results in one fusion block. All the results are acquired by partial transfer of lip.

ceptual loss to maintain the local consistency of the face identity between the source image and the result. Fig. 10 shows the effectiveness of local perceptual loss. We can observe some artifacts and color bleeding on the mouths and hair without local perceptual loss. In addition, the results with local perceptual loss are much sharper with more local details.

Network Design. Here we conduct an ablation study to evaluate the effectiveness of the network design. We aim to prove the importance of multilayer perceptron for disentanglement and the design of multiple AdaIN layers. To have a better visualization, we only transfer the lip to the source image and let the skin and eyes be the same with the source image, *i.e.*, the input is $\{y_{ref}^{lip}, y_{source}^{skin}, y_{source}^{eyes}\}$.

Qualitative comparisons are demonstrated in Fig. 9. Without MLP, the lip is entangled with skin and eyes because the initial latent space Z is not disentangled enough. By introducing an MLP to map Z to W, we can get part-specific transfer results with better disentanglement.

When only equipping with AdaIN layers in the first fusion block in MFDec, the results show much lighter transfer of the lip comparing with the reference images. Because the SCGAN only adjusts the makeup style of the results in one fusion block which is not well-refined. When emphasizing the style along the decoding process in all three fusion blocks, we can get more emphasized and confident partial makeup transfer results.

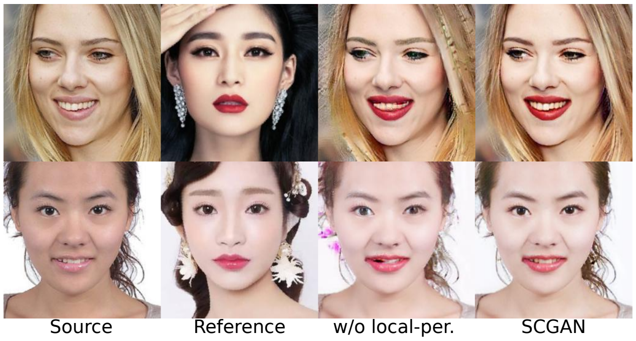


Figure 10. Ablation study of local perceptual loss.

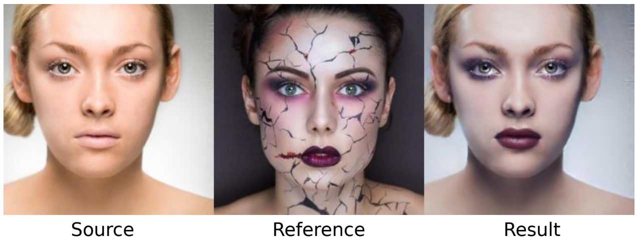


Figure 11. Limitation of our method. Although we can successfully transfer the base makeup, special effects cannot be correctly assigned.

5. Conclusion, Limitation, and Future Work

In this paper, we propose a style-codes controlled model to overcome the main challenge of spatial misalignment in makeup transfer. Unlike the previous method relies on a cumbersome and ambiguous dense correspondence between source and reference, we propose an alternative extraction-assignment solution for easing the transferring difficulty. The proposed model demonstrates a great degree of flexibility and controllability, which supports shade-controllable transfer, part-specific transfer, and makeup removal. It is invariant to spatial misalignment and rotation thanks to the style-code design. These properties perfectly match the makeup transfer application.

Although our method is flexible and accurate, a limitation is that we cannot transfer the local pattern of the facial region, as shown in Fig. 11. This may be addressed by modeling faces in the 3D domain and we leave this problem for future work.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 61972162), CCF-Tencent Open Research fund, Guangdong Provincial Key Laboratory of Cyber-Physical Systems (No. 2020B1212060069), and National & Local Joint Engineering Research Center of Intelligent Manufacturing Cyber-Physical Systems.

References

[1] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *CVPR*, pages 40–48, 2018. 1, 2, 5

[2] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. Beautyglow: On-demand makeup transfer framework with reversible generative network. In *CVPR*, pages 10042–10050, 2019. 1, 2, 5

[3] Leon Gatys, Alexander Ecker, and Matthias Bethge. A neural algorithm of artistic style. *Journal of Vision*, 16(12):326–326, 2016. 2

[4] Bruce Gooch, Greg Coombe, and Peter Shirley. Artistic vision: painterly rendering using computer vision techniques. In *Proceedings of the 2nd international symposium on Non-photorealistic animation and rendering*, pages 83–ff, 2002. 2

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 4

[6] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *ICCV*, pages 10481–10490, 2019. 1, 2, 5

[7] Dong Guo and Terence Sim. Digital face makeup by example. In *CVPR*, pages 73–79. IEEE, 2009. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[9] Aaron Hertzmann. Painterly rendering with curved brush strokes of multiple sizes. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 453–460, 1998. 2

[10] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001. 2

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 2

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 4

[13] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *CVPR*, pages 5194–5202, 2020. 1, 2, 5

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 2, 4

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 3

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[17] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NIPS*, pages 10215–10224, 2018. 2

[18] Chen Li, Kun Zhou, and Stephen Lin. Simulating makeup through physics-based manipulation of intrinsic image layers. In *CVPR*, pages 4621–4629, 2015. 2

[19] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 645–653, 2018. 1, 2, 5

[20] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *TOG*, 36(4):1–15, 2017. 5

[21] Si Liu, Xinyu Ou, Ruihe Qian, Wei Wang, and Xiaochun Cao. Makeup like a superstar: deep localized makeup transfer network. In *IJCAI*, pages 2568–2575, 2016. 2

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4, 5

[23] Yi-Zhe Song, Paul L Rosin, Peter M Hall, and John P Collomosse. Arty shapes. In *Computational aesthetics*, pages 65–72, 2008. 2

[24] Wai-Shun Tong, Chi-Keung Tang, Michael S Brown, and Ying-Qing Xu. Example-based cosmetic transfer. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 211–218. IEEE, 2007. 2

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 2

[26] Lin Xu, Yangzhou Du, and Yimin Zhang. An automatic framework for example-based virtual makeup. In *2013 IEEE International Conference on Image Processing*, pages 3206–3210. IEEE, 2013. 2

[27] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018. 3

[28] Mingtian Zhao and Song-Chun Zhu. Portrait painting using active templates. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering*, pages 117–124, 2011. 2

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 1, 2, 4, 5