BeautyREC: Robust, Efficient, and Component-Specific Makeup Transfer

Qixin Yan[1] Chunle Guo[2] Jixin Zhao[3] Yuekun Dai[3] Chen Change Loy[3] Chongyi Li[3*]

[1] WeChat, Tencent [2] TMCC, CS, Nankai University [3] S-Lab, Nanyang Technological University

qixinyan@tencent.com guochunle@nankai.edu.cn jixinzhao0101@gmail.com

YDAI005@ntu.edu.sg ccloy@ntu.edu.sg chongyi.li@ntu.edu.sg

https://li-chongyi.github.io/BeautyREC_files/

Abstract

*In this work, we propose a **R**obust, **E**fficient, and **C**omponent-specific makeup transfer method (abbreviated as **BeautyREC**). A unique departure from prior methods that leverage global attention, simply concatenate features, or implicitly manipulate features in latent space, we propose a component-specific correspondence to directly transfer the makeup style of a reference image to the corresponding components (e.g., skin, lips, eyes) of a source image, making elaborate and accurate local makeup transfer. As an auxiliary, the long-range visual dependencies of Transformer are introduced for effective global makeup transfer. Instead of the commonly used cycle structure that is complex and unstable, we employ a content consistency loss coupled with a content encoder to implement efficient single-path makeup transfer. The key insights of this study are modeling component-specific correspondence for local makeup transfer, capturing long-range dependencies for global makeup transfer, and enabling efficient makeup transfer via a single-path structure.*

*We also contribute **BeautyFace**, a makeup transfer dataset to supplement existing datasets. This dataset contains 3,000 faces, covering more diverse makeup styles, face poses, and races. Each face has annotated parsing map. Extensive experiments demonstrate the effectiveness of our method against state-of-the-art methods. Besides, our method is appealing as it is with only 1M parameters, outperforming the state-of-the-art methods (BeautyGAN: 8.43M, PSGAN: 12.62M, SCGAN: 15.30M, CPM: 9.24M, SSAT: 10.48M).*

1. Introduction

Makeup transfer is the problem of transferring the makeup style from a reference image to a source image without changing the identity and non-makeup regions of the source image. Virtual makeup applications allow people to find well-suited makeup styles online or from reference images. More and more software companies and cosmetics companies pay attention to the development of customized makeup transfer.

Transferring makeup style between a source image and a reference image is challenging. The large misalignment is easy to lead to the makeup leak. The use of real paired training data is almost impossible. The identity and non-makeup regions of source image are fragile in the process of unsupervised learning, resulting in distorted textures and artifacts in the result. Efficient makeup transfer is also demanded as transfer algorithms are usually deployed on resource-limited and real-time devices such as mobile platforms.

To deal with these challenging issues, numerous methods [1–3, 6, 10, 12, 14, 15] have been proposed in recent years. In addition to the traditional method [7] that employs layer decomposition and layer-aware makeup transfer, contemporary methods mainly adopt unsupervised GANs to learn the transfer. This is because acquiring sufficient real paired makeup/non-makeup training data is impractical. While existing methods are capable of transferring makeup, they have some inherent shortcomings. Some methods [1, 2, 12] cannot cope with spatial misalignment between the reference image and the source image well due to the simple concatenation of source and reference features. Preserving the identity and non-makeup regions of the source image after makeup transfer is crucial for a good user experience; however, some methods [2, 3, 6, 10] neglect this key fact in their designs. As a result, they cannot achieve accurate makeup transfer and even damage the non-makeup background regions. Additionally, existing methods adopt complex cycle network structures. Hence, they inevitably lead to a high memory footprint and long inference time, which constrain their practical applications. For instance, the trainable parameters of state-of-the-art methods are BeautyGAN [12]: 8.43M, PSGAN [10]: 12.62M, SCGAN [3]: 15.30M, CPM (color subnet only) [15]: 9.24M and SSAT [17]: 10.48M.

In this paper, we propose a **R**obust, **E**fficient, and

*Chongyi Li (chongyi.li@ntu.edu.sg) is the corresponding author.

Figure 1. **A set of examples of *BeautyREC***. **Left**: Our method can effectively transfer diverse makeup styles, handle different ages, and preserve the identity and non-makeup regions of the source images. **Right**: Our method can implement flexible makeup transfer, such as handling the large spatial misalignment between the reference image and the source image, achieving the makeup removal by swapping the source image and the reference image, and producing the component-specific makeup transfer (only lips).

Component-specific makeup transfer method, abbreviated as **BeautyREC**, to overcome the aforementioned issues. Instead of leveraging global attention [14], simply concatenating source features and reference features [1, 2, 12] or implicitly manipulating the component features in latent space [3], we devise a component-specific correspondence together with the corresponding component-specific discriminators to elaborately transfer the makeup styles of different components (e.g., skin, lips, eyes) in the reference image to the corresponding components of the source image. This not only avoids the artifacts induced by spatial misalignment but also preserves the non-makeup regions of the source image well.

The most related work to our component-specific correspondence is the part-specific style encoder of SCGAN [3]. Both works use the parsing maps to extract the component features. Different from SCGAN which implicitly maps the component features into an intermediate latent space and fuses them with the source features by a fusion block, our method explicitly transfers the component-specific makeup to the source image via our component-specific correspondence. Besides, unlike SCGAN which discards the spatial information of makeup features by encoding them into a style code, we use spatial information of makeup features, which benefits the transfer of spatial makeup style such as the cheek color. Note that the parsing maps are commonly used in makeup transfer methods and the current parsing map estimation and semantic segmentation methods [21] are stable in most cases. To achieve more effective global makeup transfer, we use a Transformer-based structure, as an auxiliary of the component-specific correspondence, to capture the long-range visual dependencies between the source image and reference image. Such a synergy of local and global makeup transfer cannot be achieved by previous methods.

To preserve the image content of the source image, all existing makeup transfer methods adopt complex CycleGAN [22] structures that introduce the cycle consistency loss to convert the image between the source domain and the reference domain. However, we found that a content consistency loss that constrains the content similarity between the transferred image and the source image in the feature space, coupled with a content encoder, could implement an efficient single-path makeup transfer network and preserve the content of the source image well. Consequently, the cycle structures for makeup transfer are no more required. Our model has only about 1M parameters, which outperforms state-of-the-methods by a large margin. *We wish to emphasize that it is non-trivial to make a makeup transfer model such efficient.*

In Fig. 1, we show a set of visual results of our method. As shown, the transfer of diverse makeup styles and large spatial misalignment between the reference images and the source image is made possible in our method. Moreover, our method effectively preserves the identity and non-makeup regions of the source images. Apart from the robustness, accuracy, and efficiency, our method also supports diverse applications such as makeup removal and component-specific makeup transfer. Notably, these flexible applications are implemented with a single model.

To facilitate the research on makeup transfer, we contribute a new makeup transfer dataset, **BeautyFace**, to supplement existing datasets. In comparison to existing datasets, our dataset contains more diverse makeup styles, face poses, and races. Accompanied with each face image, we also provide its parsing map.

Our main contributions are summarized as follows. **(1)** We propose a component-specific correspondence for ac-

curate component-to-component makeup transfer. **(2)** We propose a Transformer-based global makeup transfer, which models the long-range visual dependencies between the reference image and the source image. **(3)** We employ a content consistency loss coupled with a content encoder, which endows our method with an extremely lightweight makeup transfer structure.

2. Related Work

Traditional makeup transfer methods mainly employ layer decomposition [7] and face landmarks detection [19] to transfer the makeup of an example to a source image.

Recently, deep learning has been widely used in makeup transfer, especially focusing on the usage of unsupervised GANs [1–3,6,10,12,14,15,17,20]. All these methods adopt Cycle-GAN [22] structures to preserve the content of the source image. However, cycle structures need more training time and are unstable to preserve the identity and non-makeup regions of the source image. For example, BeautyGAN [12] employed cycle consistency loss, perceptual loss, adversarial loss, and makeup loss to train the makeup transfer network. To compute the makeup loss, the parsing maps of the source image and reference image are used. PSGAN [10] was proposed to improve the robustness of makeup transfer for the large pose and expression differences between the reference image and the source image. In PSGAN, the makeup of the reference image was disentangled as two spatial-aware makeup matrices that were used to modify the features of the source image for achieving the corresponding makeup style. However, it is difficult to accurately transfer the makeup style to the source image with such global attention. The parsing maps of the source image and reference image are also needed to compute the makeup loss in the PSGAN. To overcome the limitation of PSGAN that uses global attention, SCGAN [3] separately extracted the part-specific style features from the reference image and encoded them into a style-code in an intermediate latent space. The parsing maps of the reference image are needed for extracting the part-specific style features. The part-specific style code is fused with the identity code via a makeup fusion block. Although SCGAN can extract the part-specific features, the feature representations of each part are vague in latent space. Moreover, the feature fusion process of SCGAN only considers the statistic information (i.e., using AdaIN [9]) but neglects the spatial style. SOGAN [14] proposed a shadow and occlusion robust method for makeup transfer, in which the reference image and the source image are combined in the UV space. EleGANt [20] utilized high-resolution feature maps to preserve high-frequency makeup features beyond color distributions. SSAT [17] proposed a semantic-aware Transformer network and a weakly supervised semantic loss to achieve semantic correspondence.

As another line, Nguyen et al. [15] focused on both makeup color transfer and pattern transfer. For makeup color transfer, this method follows the traditional cycle structure. To implement pattern transfer, a parallel branch with the makeup color transfer branch is used to estimate the pattern mask that copes the pattern on the reference image and pastes it on the source image. Similar to the majority, the focus of our study is to transfer the makeup styles, excluding the pattern transfer. Thus, we only compare our method with the makeup color transfer branch of Nguyen et al.'s method.

3. BeautyREC

3.1. Network Structure

Overview. BeautyREC is a single-path structure, as illustrated in Fig. 2. First, the content features of the source image are extracted by the content encoder. With the use of content consistency loss in feature space, the content features are insensitive to the makeup of the source. More discussions are provided in the Ablation Study. Second, a global style encoder is used to obtain the global makeup style of the reference image while a component style encoder aims to extract the style features of different components of the reference image. Third, with the features of lips style, skin style, and eyes style of the reference image, we transfer them to the corresponding component of the source image using a component-specific correspondence. Fourth, with the global features of the reference image, the long-range visual dependencies between the reference image and the source image are modeled by the multi-head self-attention. At last, image reconstruction is employed to integrate features and produce a makeup transferred image. We provide detailed network structure and parameters in the supplementary material.

Content Encoder. In the practical applications of makeup transfer algorithms, the source image is usually covered by makeup, which increases the difficulty of transferring makeup from a reference image to the source image. This also may lead to the makeup overlay in the final result. However, this issue is commonly neglected in previous methods. Thus, they prefer the non-makeup source image in the inference process. To cope with this issue, we use a content encoder together with a content consistency loss in feature space to make the content encoder features of the source image insensitive to makeup style. The content consistency loss will be detailed in the Objective Function.

To achieve an efficient network, the content encoder contains only three Conv-IN-ReLU layers and three Resblocks, as shown in Fig. 2. Each Resblock includes two convolution layers with a residual connection. We downsample the features of the first Conv-IN-ReLU layer.

Style Encoder. We adopt the same network structure to

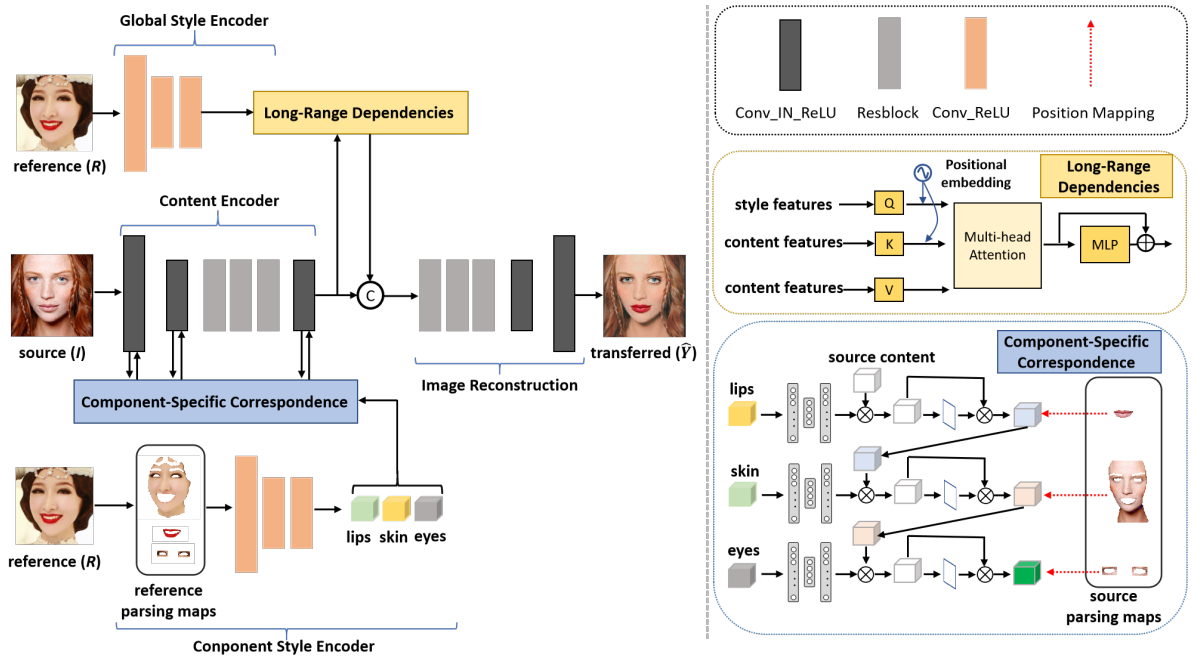


Figure 2. **Overview of BeautyREC.** It consists of a content encoder, a component style encoder, a global style encoder, a component-specific correspondence, a long-range dependency, and an image reconstruction. Note that the skip-connections between the content encoder and image reconstruction are removed in figure for brevity.

extract the component style features (i.e., component style encoder) and the global style features (i.e., global style encoder), respectively. The difference between the two style encoders is that the component style encoder separately extracts the features of different components using the corresponding parsing map, producing the component-specific features. The component style encoder also endows the flexible controllability of component-specific makeup transfer.

Specifically, we first binarize the parsing maps R_{par} of the reference image R, i.e., setting the corresponding component region (skin, lips, or eyes) to 1 and other regions to 0. Then, the corresponding component of the reference image is obtained by

$$R^{com} = R \odot \tilde{R^{com}_{par}}, com \in \{skin, lips, eyes\}, \quad (1)$$

where R^{com} represents the component com of the reference image, $\tilde{R^{com}_{par}}$ represents a binarized parsing map, and $\odot$ is the Hadamard product. R^{com} is separately fed to the three Conv-ReLU layers to achieve the corresponding component's makeup features. After the first Conv-ReLU layer, a downsampling operation is followed.

Component-Specific Correspondence. We propose a component-specific correspondence to perform accurate makeup transfer for different components, taking the statistic information and spatial information of makeup style into account in the makeup transfer process.

As shown in the bottom right corner of Fig. 2, with three sets of component-specific makeup features, the content features of the source image go through a *component-to-component* (from the reference's component to the source's component) transfer and a *component-by-component* (from lips, skin, to eyes in the source image) transfer. Following the arrows, we transferred the lips, skin, and eyes styles of the reference image one by one to the same components of the source image according to the corresponding semantic parsing map of the source image. Taking the skin style transfer as an example, the source content used here is the lips style transferred features. We first use channel attention to scale the features of content features from a statistical perspective. Then, we further process the scaled features by spatial attention. To accurately transfer the specific component's makeup style to the corresponding region of the source image, we adopt the position mapping to only change the features in the corresponding region. The process can be formulated as:

$$F^{lips}_{trans} = \text{PM}(\text{SA}(\text{CA}(F^{lips}_{style}) \otimes F_{cont})), \quad (2)$$

$$F^{skin}_{trans} = \text{PM}(\text{SA}(\text{CA}(F^{skin}_{style}) \otimes F^{lips}_{trans})), \quad (3)$$

$$F^{eyes}_{trans} = \text{PM}(\text{SA}(\text{CA}(F^{eyes}_{style}) \otimes F^{skin}_{trans})), \quad (4)$$

where F^{lips}_{tra}, F^{skin}_{trans}, and F^{eyes}_{tra} are the only lips transferred features, lips and skin transferred features, and lips, skin, and eyes transferred features. F^{lips}_{style}, F^{skin}_{style}, and F^{eyes}_{style} are the style features that refer to the lips, skin, and eyes of the

reference image. F_{cont} denotes the content features of the source image. $\otimes$ represents the pixel-wise multiplication. PM, SA, and CA represent the position mapping, spatial attention, and channel attention, respectively.

In the position mapping PM, we first compute the binarized parsing map $I_{par}^{\tilde{com}}$ of the source image I. Then, the output F_{PM}^{com} of the position mapping can be expressed as:

$$F_{PM}^{com} = F_{SA} \odot I_{par}^{\tilde{com}} \oplus F_{in} \odot (1 - I_{par}^{\tilde{com}}), \quad (5)$$

where F_{SA} denotes the output features of the corresponding spatial attention, F_{in} represents the input features of the corresponding channel attention, and $\oplus$ represents the pixel-wise addition. In this way, we only transfer the components' makeup styles of the reference image to the source image's corresponding components, thereby avoiding damaging the non-makeup regions of the source image.

In Fig. 3, we show the feature changes in the process of component-specific feature transfer. As shown, the feature of source content changes in the specific regions from lips, skin, to eyes, using the component-specific correspondence between the reference style features and the source content features, thus implementing the component-specific feature transfer. Additionally, it is insensitive to the order of feature transfer, which is discussed in the supplementary material.

Long-Range Dependencies. The component-specific correspondence may be insufficient for processing global makeup style transfer because of the inherent limitations of convolution layers such as the local modeling properties. Thus, we employ a Transformer [5,13] to exploit long-range visual dependencies between the source image and reference image.

As illustrated in Fig. 2, the basic components of our Transformer are Query (Q): style features, Key (K): content features, and Value (V): content features. It is intuitive that the style features are used as Query to model the dependencies with content features (Key). The Query is used to model the long-range visual dependencies with the Key via multi-head attention [11]. The use of multi-head attention allows our network to jointly attend the information from different representation spaces of different positions. With the obtained long-range dependencies between Query and Key, the purpose is to transfer the makeup to the Value globally. With the attention maps, the Value is weighted. The weighted Value goes through an MLP and then produces the output features. The process can be expressed as:

$$F_{gstyle} = \text{MLP}(\text{MHA}(Query, Key, Value; Pos)), \quad (6)$$

where F_{gstyle} represents the output features of global transfer, MLP is a two-layer MLP with a residual connection, MHA is a multi-head attention module with eight heads, and Pos is the sine and cosine-based position embedding.

Image Reconstruction. Based on the makeup style transferred features from the component-specific transfer and the global transfer, we employ an image reconstruction to refine the features and recover the image resolution. The image reconstruction has a symmetrical structure with the content encoder, in which the component-specific transferred features and the global features are concatenated with the corresponding decoder features.

3.2. Objective Function

Content Consistency Loss. To reduce the effect of the makeup of source image on makeup transfer performance and preserve the content of source image in the transferred result, we constrain the content consistency in feature space:

$$\mathcal{L}_{\text{cont}} = \left\| \theta_{\text{cont}}(I) - \theta_{\text{cont}}(\hat{Y}) \right\|, \quad (7)$$

where I represents the source image, $\hat{Y}$ represents the transferred result, $\mathcal{L}_{\text{cont}}$ denotes the content consistency loss, θ_{cont} is the first Conv-IN-ReLU layer of our content encoder.

Makeup Loss. Following previous methods [3, 12], our method also uses the makeup loss that consists of local histogram matching on different components of the source images and the reference image:

$$\mathcal{L}_{\text{mu}} = \|G(I,R) - \text{HM}(I,R)\|_2 + \|G(R,I) - \text{HM}(R,I)\|_2, \quad (8)$$

where $\mathcal{L}_{\text{mu}}$ denotes the makeup loss, R represents the reference image, G denotes our makeup transfer network, and $\text{HM}(\cdot)$ is the histogram matching.

Perception Loss. To preserve the perception similarity between source image and transferred result, we denote the perception loss $\mathcal{L}_{\text{per}}$ as:

$$\mathcal{L}_{\text{per}} = \left\| \theta_{\text{vgg}}(I) - \theta_{\text{vgg}}(\hat{Y}) \right\|_2, \quad (9)$$

where θ_{vgg} represents the pre-trained VGG-19 network [16]. We use the conv4 layer before the activation function.

Adversarial Loss. In addition to the global adversarial loss $\mathcal{L}_{\text{ad}}^{\text{global}}$, we also employ local adversarial losses to further enhance the significance of the local makeup style. Thus, our method is equipped with five discriminators, including a global discriminator, a skin discriminator, a lips discriminator, a left eye discriminator, and a right eye discriminator. The final adversarial loss can be expressed as:

$$\mathcal{L}_{\text{ad}} = \mathcal{L}_{\text{ad}}^{\text{global}} + \mathcal{L}_{\text{ad}}^{\text{skin}} + \mathcal{L}_{\text{ad}}^{\text{lips}} + \mathcal{L}_{\text{ad}}^{\text{leye}} + \mathcal{L}_{\text{ad}}^{\text{reye}}, \quad (10)$$

where $\mathcal{L}_{\text{ad}}^{\text{skin}}$, $\mathcal{L}_{\text{ad}}^{\text{lips}}$, $\mathcal{L}_{\text{ad}}^{\text{leye}}$, and $\mathcal{L}_{\text{ad}}^{\text{reye}}$ are the local component-specific adversarial losses.

The total loss is a combination of the above-mentioned losses, which can be expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{mu}} + \lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{ad}}\mathcal{L}_{\text{ad}}, \quad (11)$$

where λ_{per}=0.005 and λ_{ad}= 0.5 are the corresponding weights for balancing the magnitudes of losses.

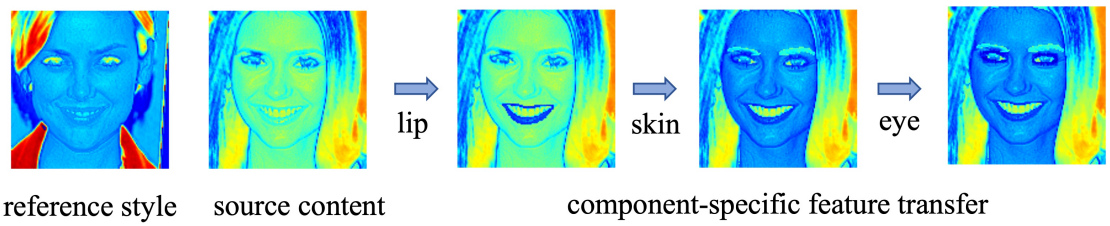


Figure 3. **Visualization of feature transfer process in our component-specific correspondence.** We normalize the values of feature maps to the range of [0,1] and average them for visualization in heatmaps.

3.3. BeautyFace

While there are some makeup datasets [10, 12], their diversity is insufficient and resolution is low (commonly 256×256). Especially, some of them were collected several years ago, thus excluding the new fashion styles. To supplement existing makeup datasets, we collect a new dataset from the Internet, named BeautyFace. It contains 3,000 high-quality face images with a higher resolution of 512×512, covering more recent makeup styles and more diverse face poses, backgrounds, expressions, races, illumination, etc. Besides, we annotate each face with parsing, which benefits more diverse applications. We show some examples of BeautyFace in Fig. 4. More results can be found in the supplementary material.

Figure 4. **Examples of BeautyFace dataset.** We show several face images (left) and the parsing maps (right).

4. Experiments

4.1. Experimental Settings

Implementations. BeautyREC is implemented with PyTorch on an NVIDIA 2080Ti GPU. We train the model with an ADAM optimizer with the fixed learning rate 1×10^{-4}. The mini-batch size is set to 1. For the discriminators used in our method, we adopt the U-Net discriminator [18] for producing accurate gradient feedback for local regions. We follow previous methods [3, 12] to train our method on MT dataset [12] that contains 3,834 images. We randomly select 100 non-makeup images and 250 makeup images for test. We also use the Wild [10] and BeautyFace for test.

Datasets	SCGAN	PSGAN	CPM	SSAT	BeautyREC
ArcFace ↑					
Wild	0.864	0.798	0.735	0.831	0.883
MT	0.857	0.829	0.767	0.892	0.878
BeautyFace	0.879	0.849	0.919	0.835	0.893
Fid ↓					
Wild	37.87	33.51	41.30	34.96	24.00
MT	70.59	47.91	52.97	37.33	38.14
BeautyFace	52.29	40.31	109.06	38.33	32.79

Table 1. **Identity preservation comparisons on the MT, Wild, and BeautyFace testing sets.**

Comparison Methods. We compare our method with several state-of-the-art makeup transfer methods: SCGAN [3], PSGAN [10], CPM [15] and SSAT [17]. We use the released code of these methods. We include only the released makeup color transfer model of CPM for fair comparisons. The pattern transfer of CPM is beyond the scope of this paper and existing makeup transfer methods. Note that the code and pre-trained models of many makeup transfer methods are not publicly available. Moreover, their training data are not clear for re-implementation. For fair comparisons, we only use the official models of different methods for experiments.

4.2. Experimental Comparisons

Visual Comparisons. We first show several representative visual comparisons in Fig. 5. As presented, the compared methods either produce unnatural makeup transfer or inaccurate transfer. For example, SCGAN, PSGAN, and CPM transfer the lipstick of the reference image to the teeth of the source image. SCGAN produces the artifacts in the regions of eyes. CPM introduces obvious artifacts in the transferred results. SSAT cannot effectively transfer the makeup from the reference images to the source images. In addition, SCGAN and SSAT cannot preserve the non-makeup regions well such as the washed-out background, and even damage the identity of the source images such as the eyes in the result of SCGAN. In comparison, our method not only effectively transfers the makeup style but also preserves the identity and non-makeup regions of the source images well. More results could be found in the supplementary material.

Identity Preservation Comparisons. To compare the per-

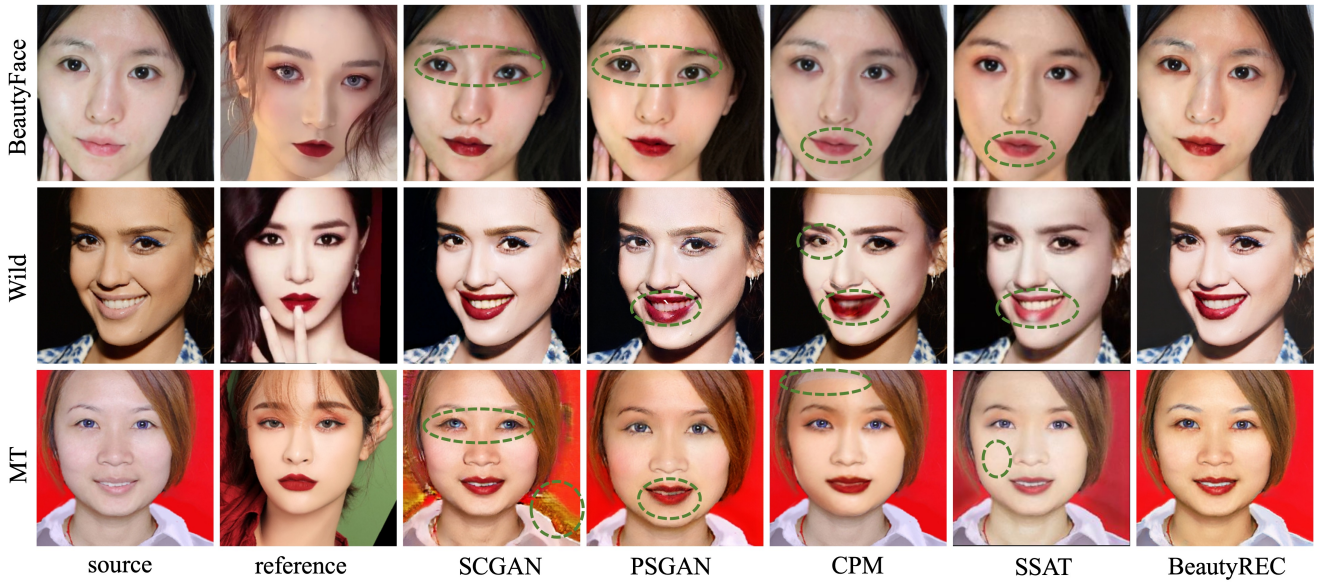


Figure 5. **Visual comparison.** Compared with the state-of-the-art methods, our method successfully transfers the makeup from the reference images to the source images. It does not introduce makeup leak and artifacts and reserves the identity and non-makeup regions of the source images well. Zoom in for best view.

formance of different methods for identity preservation of source images, we calculate the average cosine similarity of ArcFace [4] features between the faces before and after makeup transfer. We also use Fréchet Inception Distance (FID) [8] that compares the distribution of transferred images with the distribution of source images. We randomly selected 100 pairs of images from the testing set of MT [12], 100 pairs of images from Wild dataset [10], 300 pairs of images from BeautyFace dataset for test. The results of identity preservation comparison are shown in Table 1. In Table 1, our method achieves related good scores when compared with other methods. Note that the ArcFace and FID cannot accurately reflect the makeup transfer performance as they may have good scores when the transferred result keeps the same as the source image, i.e., the transfer algorithm does not work. Following previous works, we only list the scores of these two metrics as reference.

User Study. Although some methods (CPM and SSAT) can obtain relatively good performance in some identity preservation comparisons, they have poor transfer quality. Hence, we perform a user study to quantify the visual quality of the transferred results. We randomly select 10 source images and 10 reference images from Wild and 10 reference images from BeautyFace, respectively, and transfer the source images to the reference images using different methods. For each set of transferred results, we invite 20 participants to independently rank them. During ranking, these participants are trained by observing the results from 1) the makeup style similarity between the transferred result and the reference image; 2) the similarity of the identity and non-makeup regions between the transferred result and the source image; and 3) the realism of the transferred results such as artifacts and inappropriate color. We present the best-selected ratio in Table 2. Our method achieves the highest best-selected ratio, which suggests the better performance of our method for accurate makeup transfer than the compared methods.

Methods	PSGAN	CPM	SCGAN	SSAT	BeautyREC
Ratio	4.5	9.0	17.0	22.0	**47.5**

Table 2. **User study in terms of the best selected ratio (%).**

Model Size and Runtime Comparisons. We compare the model sizes and runtime in Table 3. Our method has the lowest trainable parameters and FLOPs and the fastest inference speed. The results suggest the efficiency of our method.

Methods	Parameters↓	FLOPs↓	runtime↓
SCGAN	15.30	1154.46	0.1272
PSGAN	12.62	38.82	0.1005
CPM	9.24	66.89	0.1424
SSAT	10.48	737.24	0.0681
BeautyREC	**0.99**	**12.59**	**0.0236**

Table 3. **Model size and runtime comparisons.** The trainable parameters (in M), FLOPs (in G), and runtime (in second) for processing a pair of source and reference images with a size of 256×256 are computed.

4.3. Ablation Study

We conduct ablation studies to demonstrate the effectiveness of our novel designs, including the component-specific correspondence (CSC), long-range dependencies (LRD), and content consistency loss coupled with a content encoder ($\mathcal{L}_{\text{cont}}$). We retrain the ablated models while keeping the same settings as our method, except for the ablated parts. We first conduct quantitative experiments in Table 4. As presented, the full model achieves the best identity preservation than the ablated models on Wild and MT datasets in terms of the ArcFace metric.

Datasets	w/o $\mathcal{L}_{\text{cont}}$	w/o CSC	w/o LRD	full model
Wild	0.855	0.866	0.859	**0.883**
MT	0.842	0.838	0.861	**0.878**

Table 4. **ArcFace (↑) scores of the ablated models.**

Robustness of Our BeautyREC. We provide the visual results in Fig. 6, which show the robustness of our method to the source image with or without makeup. As shown, our method can achieve the same transferred results regardless of whether the source image has makeup or not (The source B image and the source C image are covered by makeup while the source A image is not.). The results suggest that our method can eliminate the effect of the makeup on the source image, benefiting from the content encoder together with the content consistency loss in feature space.

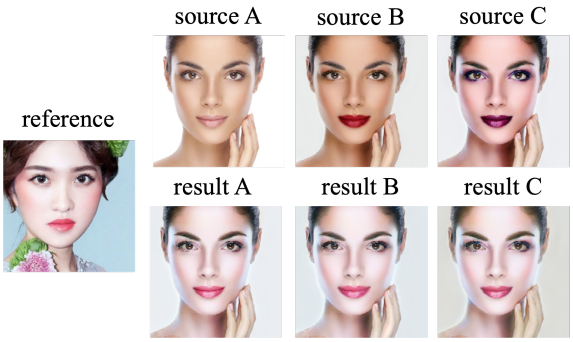


Figure 6. **Robustness of our ReautyREC.**

Effectiveness of CSC. We replace the component-specific correspondence (CSC) with simple global attention (denoted as w/o CSC) that uses channel attention and spatial attention to globally modulate the source image features by the reference image features. Such global attention is commonly used in previous makeup transfer methods. As shown in Fig. 7, the model-w/o CSC causes ambiguous makeup style transfer such as the teeth regions, and cannot transfer the local regions' makeup style well such as the eyes regions.

Effectiveness of LRD. We remove the long-range dependencies (LRD) and retrain the network. The comparison results are shown in Fig. 8. As shown, the use of long-range dependencies achieves better global makeup transfer.

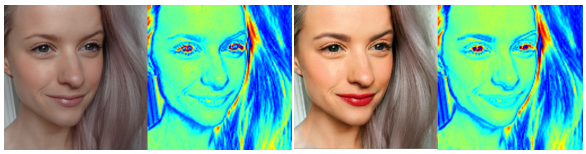
(a) source (b) reference (c) w/o CST (d) w/ CST

Figure 7. **Effect of component-specific transfer.**

(a) source (b) reference (c) w/o LRD (d) w/ LRD

Figure 8. **Effect of the long-range dependencies.**

Effectiveness of $\mathcal{L}_{\text{cont}}$. To show the effectiveness of our content consistency loss coupled with a content encoder, we separately feed the same source image with and without makeup to our method and show the features. In Fig. 9, we present the visualized features of the content encoder with the non-makeup image and makeup image as the input. With the content encoder and the proposed content consistency loss, the features of the content encoder are makeup-independent.



(a) non-makeup (b) makeup

Figure 9. **Effect of the content consistency loss coupled with a content encoder.** The features are normalized and shown in heatmaps.

5. Conclusion

In this paper, we propose a makeup transfer method to overcome the limitations of previous methods such as robustness, efficiency, and the capability of content preservation. The success of our method mainly lies in the component-specific transfer together with the global transfer and the content consistency loss. The lightweight structure and robust performance of our method outperform the state-of-the-art methods and make it suitable for practical applications. We also contribute a new makeup dataset, which facilitates the research of this research area.

Acknowledgements. This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] H. Chang, J. Lu, F. Yu, and A. Finkelstein. Pairedcycle-GAN: Asymmetric style transfer for applying and removing makeup. In *CVPR*, pages 40–48, 2018. 1, 2, 3

[2] H. J. Chen, K. M. Hui, S. Y. Wang, L. W. Tsao, H. H. Shuai, and W. H. Cheng. BeautyGlow: On-demand makeup transfer framework with reversible generative network. In *CVPR*, pages 10042–10050, 2019. 1, 2, 3

[3] H. Deng, C. Han, H. Cai, G. Han, and S. He. Spatially-invariant style-codes controlled makeup transfer. In *CVPR*, pages 6549–6557, 2021. 1, 2, 3, 5, 6

[4] J. Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. 2018. 7

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16×16 words: Transformer for image recognition at scale. *arXiv:2010.11929*, 2020. 5

[6] Q. Gu, G. Wang, M. Chiu, Y. Tai, and C. K. Tang. LADN: Local adversarial disentanling network for facial makeup and de-makeup. In *CVPR*, pages 10481–10490, 2019. 1, 3

[7] D. Guo and T. Sim. Digital face makeup by example. In *CVPR*, pages 73–79, 2009. 1, 3

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, and Bernhard Nessler. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 7

[9] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3

[10] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, J. Feng, and S. Yan. PSGAN: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *CVPR*, pages 5194–5202, 2020. 1, 3, 6, 7

[11] Jian Li, Zhaopeng Tu, Baosong Yang, Michael R. Lyu, and Tong Zhang. Multi-head attention with disagreement regularization. *arXiv preprint arXiv:1810.10183*, 2018. 5

[12] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, and W. Zhu. BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network. In *ACMMM*, pages 645–653, 2018. 1, 2, 3, 5, 6, 7

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv:2103: 14030*, 2021. 5

[14] Y. Lyu, J. Dong, B. Peng, W. Wang, and T. Tan. SOGAN: 3d-aware shadow and occlusion robust gan for makeup transfer. In *ACMMM*, 2021. 1, 2, 3

[15] T. Nguyen, A. T. Tran, and M. Hoai. Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer. In *CVPR*, pages 13305–13314, 2021. 1, 3, 6

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[17] Zhaoyang Sun, Yaxiong Chen, and Shengwu Xiong. Ssat: A symmetric semantic-aware transformer network for makeup transfer and removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2325–2334, 2022. 1, 3, 6

[18] X. Wang, L. Xie, C. Dong, and Y. Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. *arXiv:2107.10833*, 2021. 6

[19] L. Xu, Y. Du, and Y. Zhang. An automatic framework for example-based virtual makeup. In *ICIP*, pages 3206–3210, 2013. 3

[20] Chenyu Yang, Wanrong He, Yingqing Xu, and Yang Gao. Elegant: Exquisite and locally editable gan for makeup transfer. *arXiv preprint arXiv:2207.09840*, 2022. 3

[21] 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. 2

[22] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial network. In *ICCV*, pages 2223–2232, 2017. 2, 3