# RamGAN: Region Attentive Morphing GAN for Region-Level Makeup Transfer

Jianfeng Xiang[1,2,3,4], Junliang Chen[1,2,3,4], Wenshuang Liu[1,2,3,4], Xianxu Hou[1,2,3,4], and Linlin Shen[1,2,3,4]*

[1] Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University
[2] Shenzhen Institute of Artificial Intelligence & Robotics for Society
[3] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, China
[4] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, China
{xiangjianfeng2020,chenjunliang2016,liuwenshuang2018}@email.szu.edu.cn, hxianxu@gmail.com, llshen@szu.edu.cn

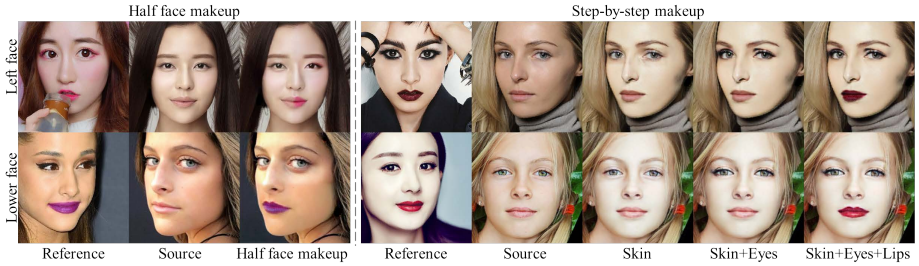



Fig. 1: Half face makeup (left) and step-by-step makeup (right).

**Abstract.** In this paper, we propose a region adaptive makeup transfer GAN, called RamGAN, for precise region-level makeup transfer. Compared to face-level transfer methods, our RamGAN uses spatial-aware Region Attentive Morphing Module (RAMM) to encode Region Attentive Matrices (RAMs) for local regions like lips, eye shadow and skin. After that, the Region Style Injection Module (RSIM) is applied to RAMs produced by RAMM to obtain two Region Makeup Tensors, $\gamma$ and $\beta$, which are subsequently added to the feature map of source image to transfer the makeup. As attention and makeup styles are calculated for each region, RamGAN can achieve better disentangled makeup transfer for different facial regions. When there are significant pose and expression variations between source and reference, RamGAN can also achieve better transfer results, due to the integration of spatial information and region-level correspondence. Experimental results are conducted on public datasets like MT, M-Wild and Makeup datasets, both visual and quantitative results and user study suggest that our approach achieves better transfer results than state-of-the-art methods like BeautyGAN, BeautyGlow, DMT, CPM and PSGAN.

**Keywords:** region makeup transfer, region attention, GAN.

* Corresponding Author

## 1 Introduction

With the development of the times, human beings, especially women are paying more and more attention to their appearance and willing to spend a lot of time and money on it. Among all facial beautification techniques, makeup is one of the most convenient and popular way, which usually applies some cosmetics like foundation, eye shadow, lipstick and so on, to generate good-looking appearance.

As facial makeup has become more and more popular, a large number of makeup transfer methods have been proposed in recent years. Makeup transfer is a computer vision task to render a non-makeup face image a makeup style without changing the face identity. Most of existing methods employ Generative Adversarial Networks (GANs) [4–6, 8, 18, 26] to learn a mapping from non-makeup face image domain to the makeup one. CycleGAN [28] adopted cycle consistency loss to learn the mapping between two domains. BeautyGAN [14] adopted the dual input/output architecture, which can perform makeup transfer and removal simultaneously. It also introduced a pixel-level histogram matching loss to improve the appearance of the lips, eye shadow and skin regions. BeautyGlow [1] used the Glow [13] framework to perform makeup transfer. LADN [9] adopted multiple and overlapping local adversarial discriminators for heavy facial makeup. DMT [25] applied two encoders to decompose the input images into identity codes and makeup codes, and produced various outputs by combining the two codes. Recently, CPM [19] successfully achieved color/pattern makeup transfer with a color/pattern transfer branch. However, most of these methods have a shortcoming, i.e., they can only work well on frontal facial images since they lack a specific module to focus on the spatial information of the images. When these methods are directly applied to the unaligned images for makeup transfer, the generated results are always far from satisfactory.

The Attentive Makeup Morphing (AMM) module proposed by PSGAN [11] tried to model how a pixel in the source is morphed from the reference image, and integrated the spatial information by including the relative positions with landmarks and the facial regions of each pixel into the attention matrix. As an extension, PSGAN++ [15] equipped an Identity Distill Network (IDNet) with the AMM module to achieve makeup transfer and removal simultaneously. However, although PSGAN and PSGAN++ can achieve makeup transfer between faces with large variations, they can not achieve accurate region-level makeup transfer, i.e., they cannot well disentangle each region when implementing partial makeup transfer.

Therefore, we propose RamGAN, which consists of two core architectures, i.e., Region Attentive Morphing Module (RAMM) and Region Style Injection Module (RSIM), for region-level makeup transfer. Fig. 1 shows four examples of region-level makeup results transferred by our approach. In the first row of the left figure, the makeup of the reference is tranferred to the left face of source and one can observe that RamGAN precisely preserves the right face of source. In the second row of the left figure, even when there is significant pose differences between source and reference, RamGAN still successfully transfers the makeup of reference to the lower part of source. In the right figue, we can observe

that RamGAN can precisely transfer the makeup for local regions like skin, eye shadow and lip.

Fig. 2 shows the main differences between our RamGAN and PSGAN. First of all, while PSGAN learns a relationship between the styles ($\gamma$ and $\beta$) of reference and source, our RamGAN does not assume such a relationship and directly learns $\gamma$ and $\beta$ applied to source. To integrate spatial constrain, the attention matrix in AMM is calculated by measuring the similarity between pixels of source and reference by weighting both the relative position to 68 landmarks and the extracted visual features. However, when faces are occluded, some of the landmarks might not be accurately detected, which will significantly affect the accuracy of attention map. Instead, the attention maps of RamGAN, Region Attentive Matrices (RAMs), are calculated for each region, based on visual features only. The makeup transfer of our approach is thus more robust against large pose differences. Based on the RAMs, two Region Makeup Tensors (RMTs) are learned to transfer the style of source face.

In addition, in order to make sure that the regions are translated separately, that is, the translation of certain region does not affect other regions, we use Region Matching Loss (RML) and Background Loss to measure the similarity between the corresponding regions of no-makeup and local/global translated images.

Our contributions are mainly summarized as follows:

- We propose a makeup transfer framework based on spatial region attention, called RamGAN, to achieve robust makeup transfer between faces with large pose variations and accurate region transfer.
- The proposed Region Attentive Morphing Module (RAMM) adaptively and separately learns the makeup information through three Region Attentive Matrices (RAMs) and successfully achieves region-level makeup transfer.
- We propose Region Style Injection Module (RSIM) to accurately transfer makeup information of reference faces to the corresponding areas of non-makeup faces.
- Experimental results quantitatively and qualitatively demonstrate that our RamGAN framework achieves the start-of-the-art performance.

## 2 Methodology

### 2.1 Problem Formulation

Let $\mathcal{X} \subset \mathbb{R}^{3\times H\times W}$ and $\mathcal{Y} \subset \mathbb{R}^{3\times H\times W}$ be the source image domain and the reference image domain. Note that the pair of makeup and non-makeup images is not available, i.e., the identities of source and reference images are different. Given a non-makeup sample $x \in \mathcal{X}$ and a makeup sample $y \in \mathcal{Y}$, the goal of the proposed RamGAN is to learn a mapping $G : x \to \tilde{y_x}$, where $\tilde{y_x} \in \mathcal{Y}$ possesses the makeup style of $y$ and the identity of $x$.

### 2.2 Network Structure

As shown in Fig. 2 **(b)**, our proposed framework mainly consists of four modules, Feature Extractor, Region Attentive Morphing Module (RAMM), Region Style Injection Module (RSIM) and Makeup Transfer Decoder.

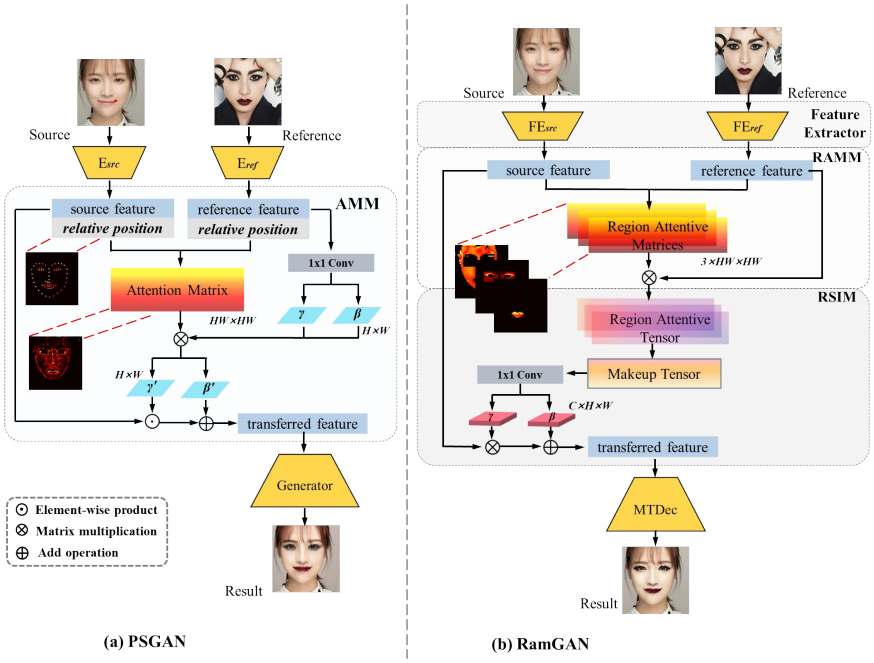



Fig. 2: An overview of PSGAN [11] **(a)** and our proposed RamGAN **(b)** .

**Feature Extractor.** As shown in Fig. 2 **(b)**, the Feature Extractor consists of two encoder-bottleneck architectures, i.e. a source image encoder $FE_x$ and a reference image encoder $FE_y$, which can extract the makeup-related features, e.g., the color of face, the size of eyes, etc. And these makeup-related features are fed to the RAMM subsequently. Note that the $FE_x$ and $FE_y$ share the same architecture, but do not share parameters. Mathematically, it is formulated as:

$$f_x = FE_x(x), \;\; f_y = FE_y(y), \tag{1}$$

where $f_x \in \mathbb{R}^{C\times H\times W}$ and $f_y \in \mathbb{R}^{C\times H\times W}$ are the source and reference feature map extracted by Feature Extractor. $C$, $H$ and $W$ are the number of channels, height and width of the feature map. $FE_x$ and $FE_y$ represent source image Feature Extractor and reference image Feature Extractor, respectively.

**Region Attentive Morphing Module.** Inspired by AMM module of PSGAN [11], we propose Region Attentive Morphing Module (RAMM) based on attention mechanism [2,3,22,24,27], which produces attention matrix for each of the facial regions like skin, lip and eye shadow. Fig. 2 shows the main differences between AMM and our RAMM. The attention matrix of AMM models the relationship between each pixel in the source with all pixels in the reference. The relative positions with 68 landmarks and the facial regions of each pixel are also considered in AMM to integrate spatial information, such that the style of reference is transferred to that of closely related pixels in source, in terms of both spatial position and visual similarity. The attention matrix is then multiplied with the style features ($\gamma$ and $\beta$) of reference and applied to the source face to

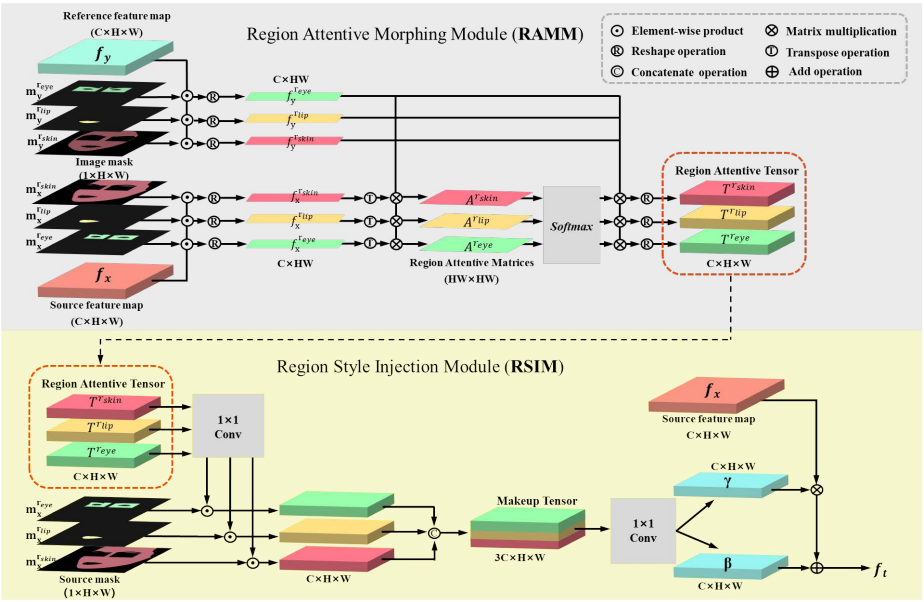



Fig. 3: Detailed architecture of the proposed RAMM and RSIM.

transfer the makeup. Different with AMM, our RAMM extracts style features and learns attention matrix for each of the facial region, the style features of reference face regions are then multiplied with the corresponding attention matrix and input to the RSIM (Region Style Injection Module) to generate the style codes ($\gamma$ and $\beta$) to transfer the makeup to source face.

As shown in the upper part of Fig. 3, our RAMM has 4 inputs, i.e., source feature map $f_x$, reference feature map $f_y$, source mask $m_x^{r_k}$ and reference mask $m_y^{r_k}$. The source feature map and reference feature map are element-wisely multiplied with corresponding facial parsing masks to get two regional feature maps.

$$f_x^{r_k} = \left(f_x \odot m_x^{r_k}\right), \; f_y^{r_k} = \left(f_y \odot m_y^{r_k}\right). \tag{2}$$

Here, $\odot$ denotes element-wise product, $f_x^{r_k}$ and $f_y^{r_k}$ indicate the source regional feature map and reference regional feature map, respectively, $m_x^{r_k}$ and $m_y^{r_k}$ represent the facial parsing mask of source and reference image, respectively. The superscript $r_k$ represents different regions of face and $k \in \{\text{skin}, \text{lip}, \text{eye shadow}\}$. Note that different definitions of regions can be used. Specially, when we perform global makeup transfer, the facial mask is defined as $m_x^{global} = m_x^{r_{skin}} + m_x^{r_{lip}} + m_x^{r_{eye}}$.

In the following branch, three Region Attentive Matrices (RAMs) are produced by multiplying two regional feature maps, which is shown in the upper part of Fig. 3. Formally, the RAMs can be expressed as:

$$A^{r_k} = \mathbf{R}(f_x^{r_k})^T \otimes \mathbf{R}(f_y^{r_k}), \tag{3}$$

where $\otimes$ and $\mathbf{R}$ denote matrix multiplication and reshape operation, respectively, and $A^{r_k}$ represents the Region Attentive Matrices for different regions.

Several differences exist between AMM of PSGAN and the proposed RAMM when calculating the attention matrix. First, both facial landmarks and facial parsing masks are required by AMM to integrate the spatial information into the attention matrix. The proposed RAMM only needs the facial masks. To integrate spatial constrain into the attention, AMM calculates the similarity between pixels of source and reference by weighting both the relative position to 68 landmarks and extracted visual features. They tested different weights and found 0.01 to be the best value, which actually emphasizes much more on the spatial positions. However, when there are significant pose variations between source and reference faces, some of the landmarks might be occluded and can't be detected. In this case, the position correspondence between source and reference might not be well established, which will significantly affect the accuracy of attention map. In contrast, our attention map is calculated for each region, based on visual features only. As faces are symmetric and the makeup styles of pixels are consistent within the same region, the makeup can still be successfully transferred to the corresponding regions when there are large pose differences between source and reference.

Fig. 4 shows the attention maps generated by PSGAN and our RamGAN for two pairs of source and references. Due to the pose differences, PSGAN wrongly matches the pixels located on the upper lip and right eyes of source face to mouth and the centers of eyes in reference face, respectively. As a result, the eye shadow and lip color of the face generated by PSGAN are significantly different with that of reference. In contrast, our RamGAN can accurately attend the makeup style to that of the same region and succesfully transfer the makeup style of each region.

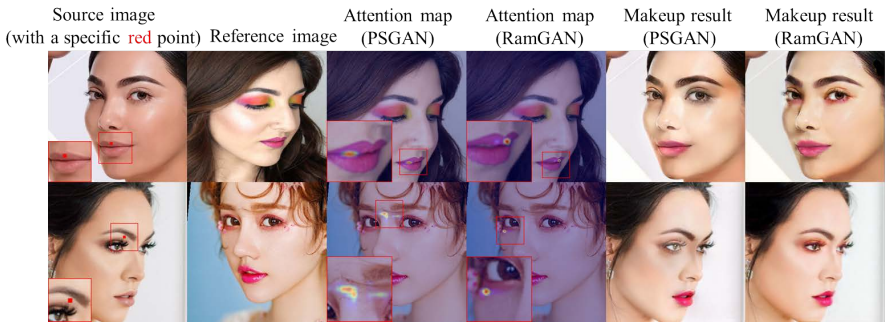



Fig. 4: The visualization of attention map on reference image. Given a specific red point in source image, we calculate the attentive values for pixels in the reference face and visualize the attention map.

After obtaining the RAMs, a softmax layer is subsequently applied to RAMs, which enables the attentive values in RAMs become more gathered [11]. Finally, the RAMM applies RAMs to the regional reference feature with matrix multiplication to produce the Region Attentive Tensor (RAT), which is shown in the red dashed box in the Fig. 3. The RAT consists of three Attentive Tensors, each of which has $C$ channels with spatial-aware attentive values for different regions. The process can be expressed as:

$$T^{r_k} = softmax\big(A^{r_k}\big) \otimes \mathbf{R}(f_y^{r_k}), \tag{4}$$

where $T^{r_k}$ represents Region Attentive Tensor for different regions, *softmax* represents softmax activation layer.

**Region Style Injection Module.** In order to accurately control the application of makeup to the target region, we introduce RSIM module. It first applies the RAT produced by RAMM to the source mask $m_x^{r_k}$ by element-wise multiplication and the output is concatenated along the channel dimension. The lower part of Fig. 3 shows the process of multiplication and concatenation. The output Makeup Tensor is fed into two $1 \times 1$ convolution layers separately to produce two Region Makeup Tensors (RMTs), $\gamma \in \mathbb{R}^{C \times H \times W}$ and $\beta \in \mathbb{R}^{C \times H \times W}$. The process can be defined as:

$$\begin{aligned} MT &= Cat\big(Conv(T^{r_k}) \odot m_x^{r_k}\big) \\ \gamma &= Conv_\gamma(MT), \beta = Conv_\beta(MT), \end{aligned} \tag{5}$$

where $MT \in \mathbb{R}^{3C \times H \times W}$ denotes Makeup Tensor, *Cat* and *Conv* represent the concatenation and $1 \times 1$ convolution, respectively. Then $\gamma$ and $\beta$ are applied to the source feature map $f_x$ to get the transferred feature map by matrix multiplication and addition. More specifically, the transferred feature map is computed by

$$f_t = \gamma f_x + \beta, \tag{6}$$

where $f_t$ represents the transferred feature map.

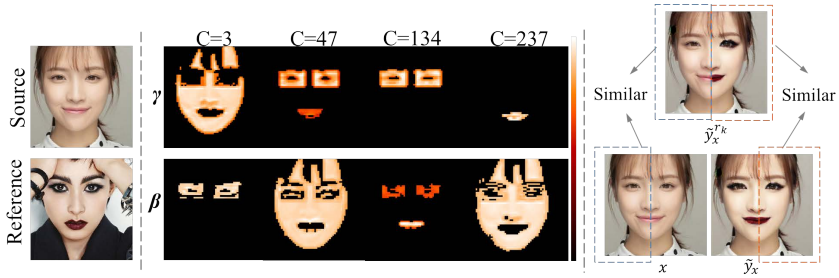
Note that the makeup matrices $\gamma'$ and $\beta'$ of PSGAN are duplicated and expanded along the channel dimension to produce the makeup tensors $\Gamma' \in \mathbb{R}^{C \times H \times W}$ and $B' \in \mathbb{R}^{C \times H \times W}$. It is unreasonable because all facial regions shared the same makeup features and thus the model has trouble in region-level makeup transfer. Different to PSGAN [11], our RMTs $\gamma \in \mathbb{R}^{C \times H \times W}$ and $\beta \in \mathbb{R}^{C \times H \times W}$ are tensors with spatial channel. We believe that makeup transfer is a region-to-region task and each channel of $\gamma$ or $\beta$ should focus on different facial regions. In Fig. 5 **(b)**, we visualize several channels of $\gamma$ and $\beta$. From the figure, we can observe that different channels of $\gamma$ or $\beta$ response to different regions, i.e. our RMTs $\gamma$ and $\beta$ contain more spatial-aware information for region-level makeup transfer.

**Makeup Transfer Decoder.** MTDec utilizes a bottleneck-decoder architecture like StarGAN [6], which is a symmetric model of Feature Extractor. The transferred feature map $f_t$ produced by RSIM is fed to the MTDec to generate the makeup result $\tilde{y}_x$, which can be expressed as:

$$\tilde{y}_x = MTDec(f_t). \tag{7}$$

### 2.3 Objective Function

**Adversarial Loss.** We employ adversarial loss to improve the quality of generated images. Given a source image domain $\mathcal{X}$ and a reference image domain $\mathcal{Y}$, we use two discriminators $D_\mathcal{X}$ and $D_\mathcal{Y}$ to distinguish generated images and real images and thus help the generator $G$ synthesize realistic outputs. Therefore, the adversarial loss of discriminators and generator can be computed by




**(a)** Source and reference **(b)** The visualization of example channels of γ and β **(c)** An example of RML

Fig. 5: **(a)** Source and reference images. **(b)** Example channels of $\gamma$ and $\beta$. **(c)** An example of Region Matching Loss.

$$\begin{aligned} L_D^{adv} = & \mathbb{E}_{x\sim\mathcal{P}_\mathcal{X}}\left[\log D_\mathcal{X}(x)\right] + \mathbb{E}_{y\sim\mathcal{P}_\mathcal{Y}}\left[\log D_\mathcal{Y}(y)\right] \\ & + \mathbb{E}_{(x,y)\sim\mathcal{P}_{(\mathcal{X},\mathcal{Y})}}\left[\log\left(1 - D_\mathcal{X}(G(y,x))\right)\right] \\ & + \mathbb{E}_{(x,y)\sim\mathcal{P}_{(\mathcal{X},\mathcal{Y})}}\left[\log\left(1 - D_\mathcal{Y}(G(x,y))\right)\right] \end{aligned} \tag{8}$$

$$\begin{aligned} L_G^{adv} = & \mathbb{E}_{(x,y)\sim\mathcal{P}_{(\mathcal{X},\mathcal{Y})}}\left[\log\left(D_\mathcal{X}(G(y,x))\right)\right] \\ & + \mathbb{E}_{(x,y)\sim\mathcal{P}_{(\mathcal{X},\mathcal{Y})}}\left[\log\left(D_\mathcal{Y}(G(x,y))\right)\right]. \end{aligned} \tag{9}$$

**Makeup Loss.** We use the makeup loss proposed in [14] to provide a coarse guidance for makeup transfer. Specifically, it employs a Histogram Matching (HM) function to adjust the color histogram distribution of the transferred image to match the reference one in each facial regions like eye shadows, lips, and facial skin. The makeup loss is a weighted sum of the regional losses

$$\mathcal{L}_G^{make} = \lambda_{lips}\mathcal{L}_{lips} + \lambda_{eyes}\mathcal{L}_{eyes} + \lambda_{skin}\mathcal{L}_{skin}, \tag{10}$$

where $\lambda_{skin}, \lambda_{eyes}$ and $\lambda_{lips}$ are tunable hyper-parameters. Specifically, each loss item is a local histogram loss, which can be written as:

$$\mathcal{L}_k = \left\|\tilde{y}_x \odot m_x^{r_k} - \mathrm{HM}\left(\tilde{y}_x \odot m_x^{r_k}, y \odot m_y^{r_k}\right)\right\|_2. \tag{11}$$

**Region Matching Loss.** As shown in Fig. 5 **(c)**, given a source image $x$, a global makeup image $\tilde{y}_x$ and a region makeup image $\tilde{y}_x^{r_k}$. We use Region Matching Loss (RML) [16,17] to measure the similarity between the $k^{th}$ regions of $\tilde{y}_x$ and $\tilde{y}_x^{r_k}$, and the similarity between other regions of $x$ and $\tilde{y}_x^{r_k}$. Then, the RML is defined as follows:

$$\mathcal{L}_G^{rm} = \left\|\tilde{y}_x^{r_k} \odot m_x^{r_k}, \tilde{y}_x \odot m_x^{r_k}\right\|_1 + \left\|\tilde{y}_x^{r_k} \odot \left(1 - m_x^{r_k}\right), x \odot \left(1 - m_x^{r_k}\right)\right\|_1, \tag{12}$$

where $1 - m_x^{r_k}$ inverts the mask to get the unrelated regions.

**Background Loss.** When performing region-level makeup transfer, we want to only change the target region, while keeping the other regions including hair, background, etc., unchanged. For this reason, we define the background loss as below

$$\mathcal{L}_G^{bg} = \left\|\tilde{y}_x \odot \left(1 - m_x^{r_k}\right), x \odot \left(1 - m_x^{r_k}\right)\right\|_1. \tag{13}$$

**Cycle Consistency Loss.** Since we are performing image-to-image translation with unpaired images, we need an additional loss to ensure that the unrelated regions in source image are not modified. Here, we introduce the cycle consistency loss proposed in [28] and define the loss function as:

$$L_G^{cyc} = \|G(G(x,y),x) - x\|_1 + \|G(G(y,x),y) - y\|_1. \tag{14}$$

**Perceptual Loss.** Perceptual loss aims to preserve the identity between source and generated images. We use the VGG-16 model [21] pre-trained on ImageNet dataset [7] to compare the activation features of source image and generated image in the hidden layer. The perceptual loss can be expressed as:

$$L_G^{per} = \|\mathscr{F}_l(G(x,y)) - \mathscr{F}_l(x)\|_2 + \|\mathscr{F}_l(G(y,x)) - \mathscr{F}_l(y)\|_2, \tag{15}$$

where $\mathscr{F}_l(\cdot)$ denotes the output of the $l^{th}$ layer of the VGG-16 model.

**Total Loss.** The total loss for discriminator and generator of our method can be expressed as:

$$\begin{aligned} L_D &= \lambda_{adv} L_D^{adv} \\ L_G &= \lambda_{adv} L_G^{adv} + \lambda_{make} L_G^{make} + \lambda_{rm} L_G^{rm} \\ &\quad + \lambda_{bg} L_G^{bg} + \lambda_{cyc} L_G^{cyc} + \lambda_{per} L_G^{per}. \end{aligned} \tag{16}$$

## 3 Experiments

### 3.1 Dataset

**Makeup Transfer dataset.** We train our RamGAN model on the Makeup Transfer (MT) dataset [14], which contains 1,115 non-makeup images and 2,719 makeup images. Most of these images consist of aligned faces with a resolution of $361 \times 361$ and provide face segmentation masks. We follow the strategy of [11] by randomly selecting 100 non-makeup and 250 makeup images as the test set and use the remaining images for training. For testing, we transfer the 100 non-makeup images with reference to each of the 250 makeup images and in total 25,000 makeup images can be generated for quality assessment.

**Makeup dataset.** LADN [9] provides Makeup dataset, which contains 333 non-makeup images, 302 makeup images and 115 extreme makeup images with great variances on makeup color, style and region coverage. We randomly select 200 non-makeup images and 200 makeup images for experiments and in total 40,000 makeup images can be generated for quality assessment.

**Makeup-Wild dataset.** Makeup-Wild [11] (M-Wild) dataset has 403 makeup images and 369 non-makeup images. Most of these images are faces with large pose variations. We randomly select 200 non-makeup images and 200 makeup images for experiments.

**CPM-Real dataset.** CPM-Real [19] dataset has 3895 real face images. Most of these images have heavy and extreme makeup, including facial gems, face paintings, hennas, and festival makeups. We select 10 non-makeup images and 10 references with light makeup for user study.

### 3.2 Implementation Details

In all experiments, we resize the images to $256 \times 256$, and use the $relu\_4\_1$ feature layer of the pre-trained VGG16 for calculating the perceptual loss. The hyper-parameters of different loss functions are set as $\lambda_{adv}$=1, $\lambda_{make}$=0.2, $\lambda_{rm}$=5, $\lambda_{bg}$=5, $\lambda_{cyc}$=10, $\lambda_{per}$=0.005. We use Adam [12] as the optimizer, the maximum epochs for model training is 50, the learning rate is 0.0002, and the batch size is 4. We implement RamGAN with Pytorch [20] and conduct all the experiments on a NVIDIA Tesla V100 GPU.

### 3.3 Qualitative Results

We compare our proposed method with the general image-to-image translation method, CycleGAN [28] and several state-of-the-art makeup transfer methods like BeautyGAN [14], BeautyGlow[5] [1], PSGAN [11], DMT [25] and CPM [19].

Fig. 6 compares the qualitative result of RamGAN with the above methods on frontal face makeup transfer. The results generated by CycleGAN have an unnatural color significantly different with the source image. Both BeautyGAN and CPM produce artifacts on the background or forehead of generated images. BeautyGlow seems to have a satisfactory result, but the color of faces, especially lips and skin, is not similar to the reference image. Comparatively, the results of PSGAN and DMT are more realistic than other methods. However, the eye shadows generated by PSGAN are all black, which are different with references. Only the results of DMT are comparable to our proposed RamGAN. However, DMT fails to achieve the makeup transfer when source and reference faces have a large difference in pose.


Fig. 6: Comparasion of frontal face makeup transfer with several state-of-the-art methods.

We also conduct an evaluation on makeup transfer between faces with large pose variations in Fig. 7. Since these methods are not equipped with a specific module to learn the spatial information, the makeup is applied randomly to the face. For example, in the first row of Fig. 7, DMT transfers the lip region into an unnatural patch. And in the second row, the makeup image generated by DMT is irrelevant to reference. In the $5^{th}$ column of Fig. 7, the faces generated by CPM are both deformed and blurry. Although the results generated by PSGAN

[5] As the source code of BeautyGlow is not available, we directly used the makeup transfer results posted on `https://github.com/BeautyGlow/BeautyGlow.github.io` for the same source and reference images for comparison.

are relatively satisfactory, the makeup styles are not accurately transferred to the appropriate regions, like eye shadow and lips, etc. As we analyzed above, the AMM module takes the relative position as the primary concern when calculating the attention matrix, which is not robust when the face is occluded.

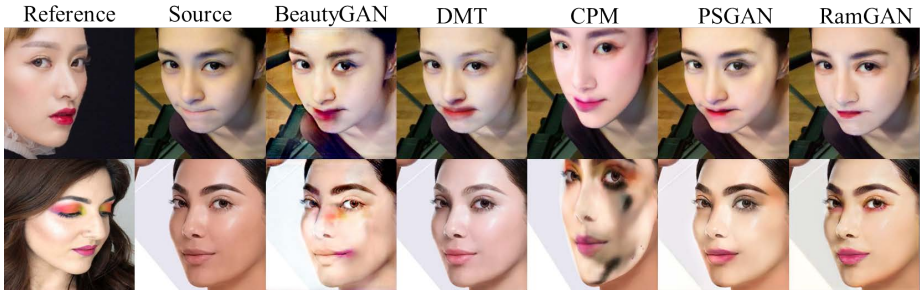



Fig. 7: Makeup transfer between faces with large pose differences.

To further illustrate that our method can not only perform global makeup transfer, but also has a strong regional controllability. We now compare step-by-step makeup transfer results with PSGAN [11]. In the first row of Fig. 8, when PSGAN performs step-by-step makeup transfer, especially changing skin color, the color of the area around eyes and lips changes simultaneously. This also proves that PSGAN cannot well disentangle each region. Though the AMM module enables pose and expression transfer, PSGAN can't perform well in region-level makeup transfer. In the second row of the figure, the proposed RamGAN succesfully achieves the step-by-step makeup transfer with a smoother and more natural transition for each region, even when there are large pose between source and reference faces.


Fig. 8: Step-by-step makeup results.

### 3.4 Quantitative Results

In this section, we demonstrate a qualitative comparison of the proposed RamGAN and other methods. We first compare Structural Similarity Index (SSIM) [23] score and Fréchet Inception Distance (FID) [10] with BeautyGAN, DMT, CPM and PSGAN on MT [14] test set, M-Wild dataset [11] and Makeup [9] dataset. Then, we conduct a user study on MT test set, M-Wild dataset, Makeup dataset and CPM-Real [19] datasest, respectively. The result images generated by all methods are aligned to the same resolution ($256 \times 256$).

**SSIM.** Structural Similarity Index (SSIM) [23] is a metric to measure the structural similarity (illumination, reflectance etc.) of two images. We use the

SSIM metric (bigger is better) to evaluate the quality of the makeup images by comparing them with source images. The average score for each method is reported in Tab. 1. The SSIM score of our method on MT test set, M-Wild dataset and Makeup dataset is **0.94**, **0.95** and **0.95**, respectively, which are higher than all other methods.

**FID.** Different from SSIM metric, Fréchet Inception Distance (FID) [10] is usually used to evaluate the quality and realness of the generated images. Therefore, we compute the FID score (smaller is better) between generated images and source images to measure our method. The result is shown in Tab. 1. We can see that our method achieves the lowest FID score among all methods.

Table 1: The SSIM/FID of different methods.

| Dataset | BeautyGAN | DMT | CPM | PSGAN | Ours |
|---|---|---|---|---|---|
| MT | 0.85/31.80 | 0.81/22.23 | 0.62/33.06 | 0.90/17.01 | **0.94/13.20** |
| M-Wild | 0.83/50.28 | 0.82/30.21 | 0.63/56.76 | 0.85/22.51 | **0.95/16.70** |
| Makeup | 0.86/38.21 | 0.90/21.16 | 0.66/43.68 | 0.90/14.83 | **0.95/10.67** |

**User Study.** To further measure the quality of images generated by our RamGAN, a user study is conducted among 65 volunteers (38 females and 27 males) aged from 20 years old to 30 years old. We randomly choose 10 non-makeup images and 10 makeup images from each of the MT test set, M-Wild dataset, Makeup dataset and CPM-Real dataset for experiments. For each of the 40 non-makeup images, the 40 makeup images are used as references and input to our RamGAN, BeautyGAN, DMT, CPM and PSGAN to generate in total 1,600 makeup transferred faces, for each model. We further divided the 1,600 makeup transfer tasks into three categories, i.e. frontal faces, faces with large pose variations and step-by-step transfer. Each volunteer was presented with the makeup faces generated by different approaches for each category of the tasks and asked to choose the best one, in terms of both image quality and identity preservation. For each category of task, four results (one from each dataset) transferred by each approach, $4 \times 5 = 20$ results, are randomly selected and shown to each of the volunteers. We in total collected 65 questionaires and each questionaire contain the best models chosen by volunteer for each category of the tasks.

Table 2: The ratio selected as best (%).

| Makeup transfer tasks | BeautyGAN | DMT | CPM | PSGAN | Ours |
|---|---|---|---|---|---|
| Frontal faces | 0.18 | 0.20 | 0.18 | 0.15 | **0.29** |
| Faces with large pose variations | 0.03 | 0.17 | 0.03 | 0.10 | **0.67** |
| Step-by-step | — | — | 0.08 | 0.15 | **0.77** |

Tab. 2 shows the ratio of each model chosen by volunteers and it shows that our RamGAN is the most frequently chosen model across all of the different tasks. Especially for makeup transfer across large pose variations and stey-by-step transfer, the ratio of our approach chosen by volunteers is significantly higher than other competing approaches. As the stey-by-step makeup transfer resuts of BeautyGAN and DMT are far from satisfactory, we don't include them into the questionaire for the step-by-step task.

### 3.5 Ablation Studies

We now test the effectiveness of the proposed RAMM and Region Matching Loss (RML). As presented in previous sections, our RAMM mainly enables region-level transfer and RML further reduces the entanglement of different regions to the target region. Fig. 9. shows the results of two examples for our RamGAN with and without the proposed RAMM and RML. As shown in the $4^{th}$ column, the forehead of the lady's face in the first row and the mouth region in the second row are also changed when RamGAN without RML is trying to transfer the makeup of the mouth and eye regions, respectively. Instead, the target regions transferred by RamGAN with RML are much more precise and other regions are preserved much better. In the $5^{th}$ column, there are obvious transition boundaries in the makeup faces transferred by RamGAN without RAM, which clearly justifies the usefulness of the proposed module.

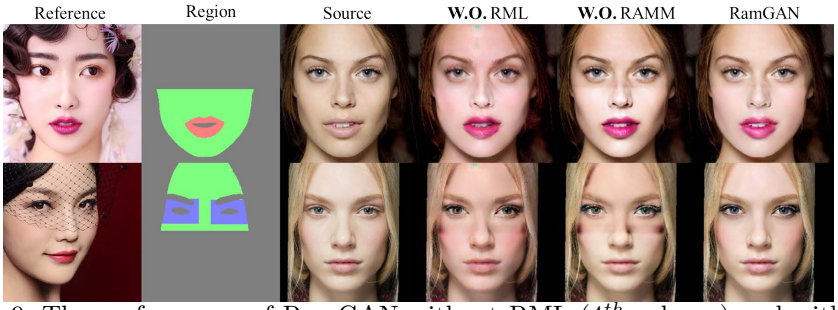



Fig. 9: The performance of RamGAN without RML ($4^{th}$ column) and without RAMM ($5^{th}$ column).

We now show the quantitative results of SSIM and FID for RamGAN without RAMM and RML in Tab. 3. One can observe from the table that RamGAN equipped with the two modules achieves much better results.

Table 3: The SSIM and FID of ablation study.

| Metric | Dataset | **W.O.** RAMM | **W.O.** RML | RamGAN |
|---|---|---|---|---|
| SSIM | MT | 0.72 | 0.61 | **0.94** |
| | M-Wild | 0.73 | 0.63 | **0.95** |
| | Makeup | 0.71 | 0.69 | **0.95** |
| FID | MT | 30.21 | 45.75 | **13.20** |
| | M-Wild | 35.73 | 50.83 | **16.70** |
| | Makeup | 32.36 | 16.19 | **10.67** |

### 3.6 More Visual Results

Based on our RAMM and RSIM, we can actually perform mixed style transfer by transferring different regions to the styles of different faces. Given three reference images, $y_0$, $y_1$, $y_2 \in \mathcal{Y}$, we can obtain the corresponding makeup-related features $f_{y_0}$, $f_{y_1}$ and $f_{y_2}$ extracted by Feature Extractor, respectively. Based on the facial region masks, $m_{y_0}^{r_{skin}}$, $m_{y_1}^{r_{lip}}$ and $m_{y_2}^{r_{eye}}$, we can obtain the corresponding regional

feature maps $f_{y_0}^{r_{skin}}$, $f_{y_1}^{r_{lip}}$ and $f_{y_2}^{r_{eye}}$ with Eq. (2). Thereafter, different regions of a source face can be transferred to the styles of corresponding regions encoded in the three different feature maps. For example, in middle of the $2^{nd}$ row of Fig. 10, the skin, lips and eye shadow of the source image are transferred to the styles of corresponding regions of the three references shown in the first row, respectively. The last facial image in the $2^{nd}$ row shows the results of mixed transfer by integrating the styles of three regions of different reference faces, i.e. the skin, lips and eye shadow of the face are similar to the styles of the three reference faces shown in the first row, respectively. More results of interpolation between difference references can be found in the supplementary.

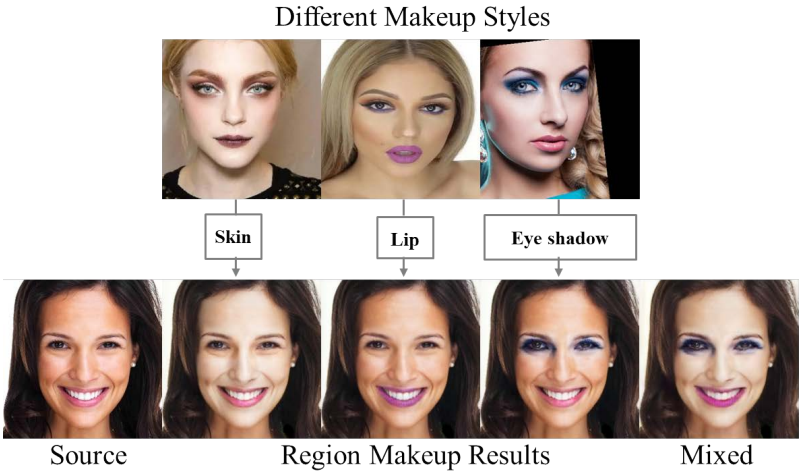



Fig. 10: The mixed trasfer of different makeup styles. First rows are different styles. Second rows are source image, region makeup transfer results (skin, lips and eye shadow), and mixed result.

## 4 Conclusions

In this paper, we discuss the makeup transfer task, which aims to render a non-makeup face image a makeup style without changing the face identity. We propose a region attentive morphing generative adversarial network (RamGAN) for facial makeup transfer. Our RamGAN can achieve state-of-the-art results, which performs region-level makeup transfer and makeup transfer between faces with large pose variations. Extensive experiments on various datasets further demonstrate that our method significantly outperforms the latest makeup transfer approaches e.g. BeautyGAN, BeautyGlow, DMT, CPM and PSGAN. Moreover, our method has a great advantage in precise region control. Therefore, we believe that our method can be applied to other regional image-to-image translation task.

## 5 Acknowledgements

This research was supported by National Natural Science Foundation of China under grant no. 91959108, and Guangdong Basic and Applied Basic Research Foundation under Grant no. 2020A1515111199 and 2022A1515011018.

## References


1. Chen, H.J., Hui, K.M., Wang, S.Y., Tsao, L.W., Shuai, H.H., Cheng, W.H.: Beautyglow: On-demand makeup transfer framework with reversible generative network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10042–10050 (2019)
2. Chen, J., Lu, W., Shen, L.: Selective multi-scale learning for object detection. In: International Conference on Artificial Neural Networks. pp. 3–14. Springer (2021)
3. Chen, J., Zhao, X., Shen, L.: Delving into the scale variance problem in object detection. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 902–909. IEEE (2021)
4. Chen, W., Shen, L., Lai, Z.: Introspective gan for meshface recognition. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 3472–3476. IEEE (2019)
5. Chen, W., Xie, X., Jia, X., Shen, L.: Texture deformation based generative adversarial networks for multi-domain face editing. In: Pacific Rim International Conference on Artificial Intelligence. pp. 257–269. Springer (2019)
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: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
9. Gu, Q., Wang, G., Chiu, M.T., Tai, Y.W., Tang, C.K.: Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10481–10490 (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. Advances in neural information processing systems **30** (2017)
11. Jiang, W., Liu, S., Gao, C., Cao, J., He, R., Feng, J., Yan, S.: Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5194–5202 (2020)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. arXiv preprint arXiv:1807.03039 (2018)
14. Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., Lin, L.: Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 645–653 (2018)
15. Liu, S., Jiang, W., Gao, C., He, R., Feng, J., Li, B., Yan, S.: Psgan++: Robust detail-preserving makeup transfer and removal. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
16. Liu, W., Chen, W., Shen, L.: Translate the facial regions you like using region-wise normalization. arXiv preprint arXiv:2007.14615 (2020)

17. Liu, W., Chen, W., Yang, Z., Shen, L.: Translate the facial regions you like using self-adaptive region translation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2180–2188 (2021)
18. Liu, W., Chen, W., Zhu, Y., Shen, L.: Satgan: Augmenting age biased dataset for cross-age face recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 1368–1375. IEEE (2021)
19. Nguyen, T., Tran, A.T., Hoai, M.: Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13305–13314 (2021)
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
24. Xie, J., Luo, C., Zhu, X., Jin, Z., Lu, W., Shen, L.: Online refinement of low-level feature based activation map for weakly supervised object localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 132–141 (2021)
25. Zhang, H., Chen, W., He, H., Jin, Y.: Disentangled makeup transfer with generative adversarial network. arXiv preprint arXiv:1907.01144 (2019)
26. Zhang, X., Zhu, Y., Chen, W., Liu, W., Shen, L.: Gated switchgan for multi-domain facial image translation. IEEE Transactions on Multimedia (2021)
27. Zhao, X., Chen, J., Liu, M., Ye, K., Shen, L.: Multi-scale attention-based feature pyramid networks for object detection. In: International Conference on Image and Graphics. pp. 405–417. Springer (2021)
28. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)