Real-World Automatic Makeup via Identity Preservation Makeup Net

Zhikun Huang[1*] , **Zhedong Zheng**[2,4] , **Chenggang Yan**[1†] , **Hongtao Xie**[3] ,
Yaoqi Sun[1] , **Jianzhong Wang**[1] , **Jiyong Zhang**[1]
[1]Hangzhou Dianzi University
[2]University of Technology Sydney
[3]University of Science and Technology of China
[4]Baidu Research
{hzk,cgyan,syq,wangjz,jyzhang}@hdu.edu.cn, zhedong.zheng@student.uts.edu.au, htxie@ustc.edu.cn

Abstract

This paper focuses on the real-world automatic makeup problem. Given one non-makeup target image and one reference image, the automatic makeup is to generate one face image, which maintains the original identity with the makeup style in the reference image. In the real-world scenario, face makeup task demands a robust system against the environmental variants. The two main challenges in real-world face makeup could be summarized as follow: first, the background in real-world images is complicated. The previous methods are prone to change the style of background as well; second, the foreground faces are also easy to be affected. For instance, the "heavy" makeup may lose the discriminative information of the original identity. To address these two challenges, we introduce a new makeup model, called **I**dentity **P**reservation **M**akeup Net (**IPM-Net**), which preserves not only the background but the critical patterns of the original identity. Specifically, we disentangle the face images to two different information codes, *i.e.*, identity content code and makeup style code. When inference, we only need to change the makeup style code to generate various makeup images of the target person. In the experiment, we show the proposed method achieves not only better accuracy in both realism (FID) and diversity (LPIPS) in the test set, but also works well on the real-world images collected from the Internet.

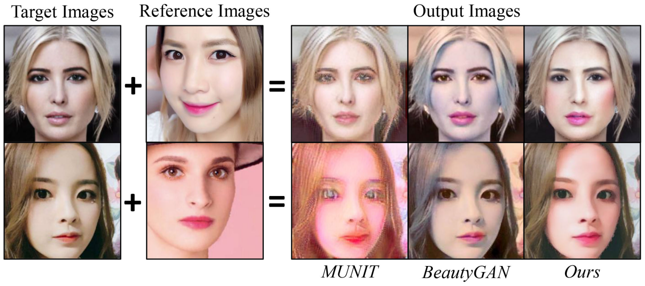


Figure 1: The results of compared methods, *i.e.*, MUNIT [Huang *et al.*, 2018] and BeautyGAN [Li *et al.*, 2018], as well as our results on real-world images. The target images are collected from the Internet, and the models have not seen them before. Both MUNIT and BeautyGAN suffer from complicated backgrounds and over-makeup. Our proposed method yields high-fidelity face images.

1 Introduction

In recent years, an increasing number of people begin to share their photos on social networks, such as Facebook, to express their feelings [Senft and Baym, 2015; Sorokowski *et al.*, 2015]. People usually want to upload good-looking selfies to social media, because beautiful photos could catch other eyes easily. Although the endless stream of cosmetics give people more alternatives for makeup, people are eager to obtain a beautiful picture more effective and efficient. Automatic makeup is desirable. Given one non-makeup target image and one reference image, the automatic makeup is to generate one face image, which maintains the original identity with the makeup style in the reference images. However, automatic makeup meets several challenges in real-world practice. As shown in Fig. 1, we observe that there are two main problems in the existing methods. First, when the background is complicated, the previous methods may modify the background styles. This challenge usually results in the local inconsistency, which leads to unrealistic artifacts. Second, the makeup degree is hard to control. The over-makeup may overwrite critical facial patterns, which loses the discriminative information of the original identity. It is undesirable to share one selfie that others could not recognize you.

To address the above two problems, we propose a new makeup model called **I**dentity **P**reservation **M**akeup Net (**IPM-Net**), which retains not only the background but the critical information of the target identity. Specifically, we first deploy the traditional image processing method to preprocess the input images. The pre-processing method predicts the segmentation maps and provides the geometric information of face images. Given the segmentation maps, we generate the mask and extract the fine-grained texture for every face. Given the face masks and textures, we keep the identity contents, which contain the non-makeup areas, such as wall, eyes, and hair. We further disentangle the images to two codes, *i.e.*, **identity content code** and **makeup style**

[*] Work done during a visit at University of Technology Sydney.
[†] Contact Author

Identity Content Space	Makeup Style Space
body, background*, facial structure and texture, ears, eyes, neck, etc.	makeup foundation, eyebrow, eye shadow, lipstick, etc.

Table 1: **Identity content space:** personal identity in the image and other details that should be kept. **Makeup style space:** makeup styles of the face. *: background is expected to be maintained as much as possible, to avoid unrealistic artifacts. We, therefore, regard the background as identity content space.

code. The identity content code contains personal identifying information in the face image and other details that should be kept, while the makeup style code learns the makeup style of the reference face. We note that background is expected to be preserved as much as possible, to avoid unrealistic artifacts. We, therefore, regard the background as the identity content space. More details are provided in Table 1. Finally, we reconstruct a "new" face image by exchanging the makeup style codes between reference images and target images. In the inference phase, we could generate different makeup images by changing different makeup style codes. Besides, we notice that almost existing studies do not have a reliable quantitative metric. Except for the conventional qualitative results, *e.g.*, displaying the makeup faces and conducting user study [Li *et al.*, 2018; Zhang *et al.*, 2019], we introduce the Fréchet Inception Distance (FID) [Heusel *et al.*, 2017] for image realism evaluation and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018] for diversity. We hope that these two metrics could help the community to evaluate the automatic markup models. In summary, our contribution is three-fold:

- We propose a new automatic makeup model, called **I**dentity **P**reservation **M**akeup Net (IPM-Net), to address the two problems in the real-world automatic makeup tasks. Our method effectively transfers the makeup style of the reference image to the target image, while preserving the background as well as the critical patterns of the original identity.
- The proposed method enables the controllable makeup transfer within a short inference time and could handle 16 face images of 256 × 256 per second. It makes our model reach one step closer to real-world practice.
- As a minor contribution, we introduce the FID and SSIM for the makeup realism and diversity evaluation. Our method outperforms other competitive methods in both qualitative and quantitative results.

2 Related Works

2.1 Hand-crafted Makeup

In the previous research, many methods provided technical and theoretical support for automatic makeup transfer. [Shashua and Riklin-Raviv, 2001] provide an approach to transfer shading from samples based on color ratio. [Leyvand *et al.*, 2006] take into account the geometric alterations of the contours of facial features to produce a more pleasant face image. [Tong *et al.*, 2007] describe an image-based procedure to achieve makeup style transfer by example. [Guo and Sim, 2009] decompose target image and reference image into three different layers containing different information. They further transfer information from each layer of the reference image to the corresponding layer of the target image. [Li *et al.*, 2015] build several physics-based mapping models for corresponding separated intrinsic image layers to complete the automatic makeup.

2.2 Deeply-learned Makeup

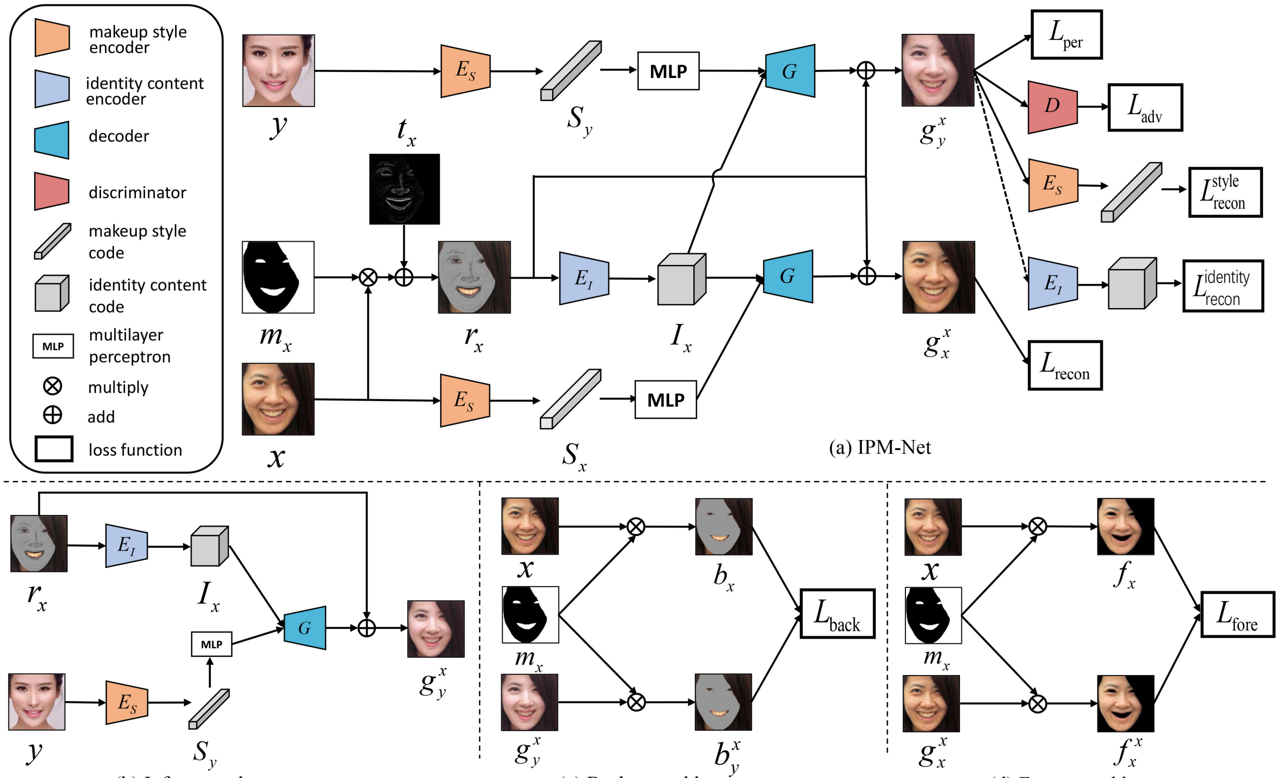
Recent years, different types of neural networks are adapted to extract deep cosmetic features and complete the complex mapping for makeup. [Liu *et al.*, 2016] parse the target image and the reference image to facilitate the makeup transfer between the corresponding regions individually. Most recent researches [Li *et al.*, 2018; Chang *et al.*, 2018; Zhang *et al.*, 2019; Jiang *et al.*, 2019; Chen *et al.*, 2019] achieve automatic makeup transfer by leveraging generation adversarial networks [Goodfellow *et al.*, 2014]. For instance, CycleGAN [Zhu *et al.*, 2017] has been exploited in many previous works [Li *et al.*, 2018; Chang *et al.*, 2018; Chen *et al.*, 2019] to learn from makeup images and non-makeup images. [Zhang *et al.*, 2019] disentangle the images into personal identity and makeup style, combining the personal identity from the target image and the makeup style from reference images to reconstruct a new image. Besides, different attention mechanisms are applied by [Li *et al.*, 2018; Zhang *et al.*, 2019; Jiang *et al.*, 2019; Jin *et al.*, 2019; Chen *et al.*, 2019] to make the networks pay more attention to the face area in images. In a similar spirit, some works [Li *et al.*, 2018; Zhang *et al.*, 2019] explicitly utilize segmentation maps to process different makeup areas separately. Furthermore, [Jiang *et al.*, 2019] introduce an attention makeup morphing module to specify the pixel in the source image that should be morphed from the reference image. [Jin *et al.*, 2019] divide images into three different layers and then apply facial landmarks to process them separately.

However, previous methods are prone to change the style of the background while transferring the makeup styles. Besides, the discriminative information of the original identity may be lost if the reference makeup style is a "heavy" makeup. In contrast, our IPM-Net focuses on preserving not only the background content but also the essential identity information. For instance, the identity content information, such as eyes, ears and the neck, should not be modified. We keep the original style of identity content. The generated face images, therefore, is closer to the real makeup images.

3 Methodology

3.1 Images Pre-processing

We note that the automatic makeup aims at only changing several parts of the target image while keeping most identity content information. We, therefore, propose to disentangle the facial images to two spaces first (see Table 1). Identity content space contains the personal identity information in the image and other details that should be kept. Makeup style space is to learn makeup styles on the face. To help the two



(b) Inference phase (c) Background loss (d) Foreground loss

Figure 2: A schematic overview of IPM-Net. (a) IPM-Net aims to generate a new image that contains the makeup style of the reference image y and the identity information of the target image x. m_x and t_x are the mask and facial texture of x, respectively. Using m_x, t_x and x, we can obtain the identity content input image r_x via the pre-processing approach described in Section 3.1. The makeup style encoder E_S extracts makeup style codes from x and y, while the identity content encoder E_I extract the identity content code of x from r_x. The decoder G fuses identity content code and makeup style code to generate a new image g_y^x. Specifically, (b) E_I encode the identity content of r_x into I_x and E_S encode the makeup style of y into S_y. G decode I_x and S_y into a new image. Furthermore, the residual information from r_x and the output of G add up to the expected generated image g_y^x. Besides, we can reconstruct the target image x using the generative module. (c) Background loss minimizes the gap between the backgrounds of the target image x and the new image g_y^x to retain background details. (d) Foreground loss is committed to ensuring that the makeup styles of reconstructed image g_x^x and target image x are the same.

different feature extraction, we pre-process the input images. Specifically, we customize a mask and a fine-grained texture for each face. The pre-processing results are shown in Fig. 3. Given one input image x, we generate the identity content input image r_x in four steps: (1) We first acquire the face mask m_x via the face parsing algorithms [Yu *et al.*, 2018]. (2) To preserve the background, we multiply the target image x and the corresponding mask m_x, which results in the background image b_x. We further set the makeup area to gray, but b_x loses the face texture information. (3) Therefore, we extract the texture of makeup areas using a differential filter [Pu *et al.*, 2008], which preserves only the necessary textures and filters out the noisy signals. Besides, we increase the weights of the texture of facial features to highlight the texture of facial features. (4) Finally, the identity content input image r_x is generated by adding the texture t_x and the background b_x.

3.2 Framework

Formulation. We denote the non-makeup images and the makeup images as X and $Y(X, Y \subset \mathbb{R}^{H\times W\times 3})$, respectively. Let one target image $x \in X$ and one reference image $y \in Y$. Given the target image x and the corresponding mask m_x and texture t_x, an identity content input image r_x is generated by the pre-processing steps described in Section 3.1. As shown in Fig. 2 (a), an identity content encoder E_I and a makeup style encoder E_S are introduced to disentangle the face image into two different codes: the target image is decomposed into identity content $I_x = E_I(r_x)$ and original makeup style code $S_x = E_S(x)$, while the reference image provides the makeup style $S_y = E_S(y)$. In order to blend I_x and S_y, the decoder G uses a multilayer perceptron (MLP) to produce a set of AdaIN [Huang and Belongie, 2017] parameters from S_y. Meanwhile, in the spirit of the previous super-resolution works [Lai *et al.*, 2018], we adopt the residual connection to keep the identity content and background of x unchanged:

$$g_y^x = G(I_x, S_y) + r_x, \tag{1}$$

where g_y^x denotes the synthetic image, which contains the identity content I_x of x and makeup style S_y of y. The inference phase is illustrated in Fig. 2 (b).

Content Maintenance. VggFace2 [Cao *et al.*, 2018] is a widely-used face recognition dataset, which contains a large amount of face data. We deploy the ResNet-50 [He *et al.*,

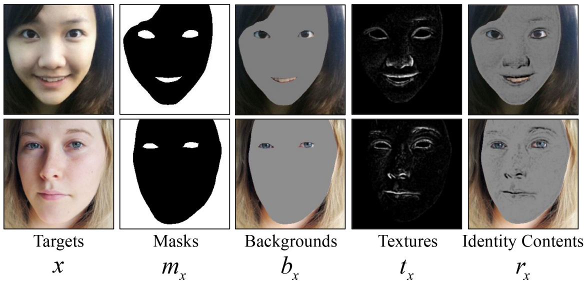


Figure 3: Description of images pre-processing. In order to understand the two spaces described in Table 1, we show our pre-processing step by step in this figure: (1) The mask m_x is customized for the face in target image using segmentation map. (2) Then, given the target image x and the mask m_x, we can gray out the makeup area in the target image and get the background b_x. (3) Also, we can retain the makeup areas and remove background, *i.e.*, leave the foreground. Using a differential filter, we extract the texture t_x of makeup areas. (4) Finally, the identity content r_x is obtained by adding texture t_x to background b_x.

2016] model trained on the VggFace2 to extract the high-level features of x and g_y^x, and then using ℓ_1 loss to ensure that they represent the same identity:

$$L_{\text{per}} = \mathbb{E}[\left\|R(x) - R(g_y^x)\right\|_1], \tag{2}$$

where $R(\cdot)$ denotes the output of ResNet-50. Comparing to the general model, such as InceptionNetV3 [Szegedy *et al.*, 2016] trained on ImageNet [Deng *et al.*, 2009], the model trained in face recognition, could better reflect the discrepancy between different identities. The weight of the ResNet-50 is fixed in the entire training process. So when we minimize the loss L_{per}, the personal identity of x is kept well.

Besides, to make the generated images more realistic and natural, the preservation of the background details is also a key to the high-fidelity generated images. As shown in Fig. 2(c), we introduce a simple but effective background loss to keep the details of the background:

$$L_{\text{back}} = \mathbb{E}[\left\|b_x - b_y^x\right\|_1], \tag{3}$$

where b_x and b_x^y are the background of the target image x and generated image g_y^x. The two backgrounds b_x and b_x^y are obtained by multiplying the original mask m_x with x and g_y^x, respectively.

Self-identity Generation. As shown in Fig. 2(a), we also reconstruct the target image x by changing the reference makeup style code S_y to original makeup style code S_x, *i.e.*, $g_x^x = G(I_x, S_x)$. The pixel-wise ℓ_1 loss is applied to help the self-reconstruction:

$$L_{\text{recon}} = \mathbb{E}[\|x - g_x^x\|_1]. \tag{4}$$

The reconstruction loss plays an important role in regularizing the generative module. However, the pixel-wise ℓ_1 loss is hard to handle every detail in the image, for example, the color of the lips or eye shadow transform slightly. So we further introduce a foreground loss shown in Fig. 2(d), which ignores the identity content spaces and focuses on the makeup style. To minimize the gap between the foreground f_x of the target image x and the foreground f_x^x of the reconstructed image g_x^x, the foreground loss could be formulated as:

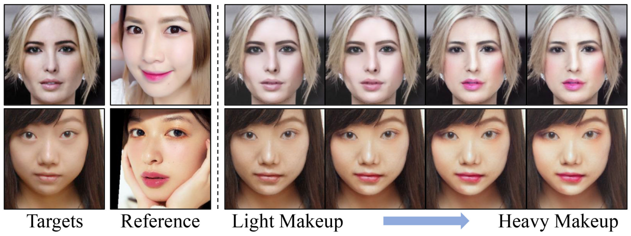


Figure 4: Controllable automatic makeup transfer results. Our model allows users to adjust different levels of makeup transfer, that reaches one step closer to the real-world practice. The target image on the first row is obtained via the web with light makeup. Another target image and two reference images come from the Makeup Transfer dataset without makeup. To allow the makeup transfer degree controllable, we deploy a new reference makeup style S_{new} as a weighted sum of the reference makeup styles S_y and the original makeup style S_x. The generated results are sorted from left to right according to the makeup transfer level from light to heavy.

$$L_{\text{fore}} = \mathbb{E}[\|f_x - f_x^x\|_1]. \tag{5}$$

Makeup Generation. As we described above, we could generate a new image g_y^x using two latent codes of two input images x and y. To preserve the identity content information I_x from target image x and the makeup style information S_y from reference image y, we introdcue two ℓ_1 losses, *i.e.*, $L_{\text{recon}}^{\text{identity}}$ and $L_{\text{recon}}^{\text{style}}$, which could be formulated as:

$$L_{\text{recon}}^{\text{identity}} = \mathbb{E}[\left\|I_x - E_I(g_y^x \cdot m_x + t_x)\right\|_1], \tag{6}$$

$$L_{\text{recon}}^{\text{style}} = \mathbb{E}[\left\|S_y - E_S(g_y^x)\right\|_1], \tag{7}$$

where $\cdot$ denotes the pixel-wise multiplication. In addition, we adopt the adversarial loss to make the generated images become indistinguishable from real images:

$$L_{\text{adv}} = \mathbb{E}[logD(x) + log(1 - D(G(I_x, S_y)))], \tag{8}$$

where the discriminator D targets to distinguish real faces from fake ones. We provide the detailed structure of E_S, E_I, G and D in Section 4.1.

4 Experiment

In this section, we first introduce the implementation details, including the dataset, the network structure, and compared approaches (See Section 4.1). After that, we report the qualitative results to verify the effectiveness of the proposed automatic markup network. Besides the makeup samples, we also show the controllable automatic makeup transfer results to manipulate the makeup levels. In Section 4.3, we introduce the two new evaluation metrics, *i.e.*, the Fréchet Inception Distance (FID) [Heusel *et al.*, 2017] and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018] to compare the proposed method with other competitive methods quantitatively. Finally, we provide a discussion on the computational efficiency of the whole markup pipeline.

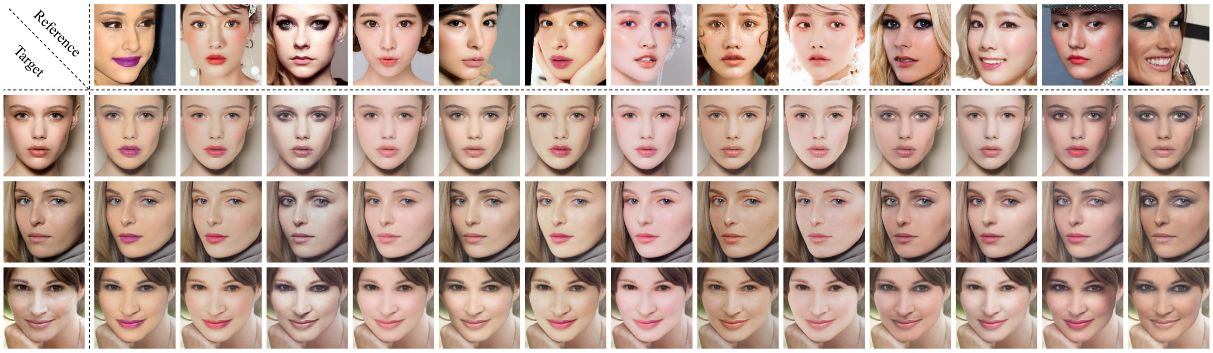


Figure 5: Automatic makeup results. Three targets images with different backgrounds and different skin color are shown in the first column. Thirteen reference images with different makeup styles and poses are shown in the first row. The synthetic results are shown in the lower right, and each row and column corresponds to different identity content and makeup style. Note that the synthetic results obtain the discriminative makeup style in the reference images, *e.g.*, the color of lipstick and eye shadow, when the identity information of target images remains.

4.1 Implementation Details

Dataset. We train and test our model on the widely-used Makeup Transfer dataset [Li *et al.*, 2018]. The Makeup Transfer dataset contains two sets of female face images with 2,719 makeup images and 1,115 non-makeup images. The segmentation map of every face is provided. In addition, we collect extra test images of celebrities from the Internet to verify the scalability of our model in the real-world scenario. As for the extra test data collected from Internet, we acquire their masks via face parsing algorithms [Yu *et al.*, 2018].

Network Structure. Our IPM-Net is implemented in Pytorch [Paszke *et al.*, 2017]. We also use PaddlePaddle to implement our method and achieve similar performances. All our experiments are conducted on one NVIDIA GTX 2080Ti GPU. We apply two kinds of basic blocks, *i.e.*, ConvBlock and ResBlock. ConvBlock contains convolution, batch normalization [Ioffe and Szegedy, 2015], and ReLU activation layers. ResBlock contains two ConvBlock, but we remove the last activation layer. Our network structure in Fig. 2(a) is built based on these two blocks: (1) E_I consists of three ConvBlocks and a ResBlock to output the identity content code I_x in 256×64×64. (2) E_S uses a combination of three ConvBlocks and three ResBlocks, while an average pooling layer added at the end. Both the target image and the reference image share the E_S, and each makeup style is represented by a 128-dim vector. (3) G adopts the ConvBlocks and utilizes AdaIN [Huang and Belongie, 2017] to fuse identity content and makeup style. An up-sample layer is further leveraged to re-scale the image as the input image shape. In addition, r_x is added to the image generated by G as the residual connection. (4) D follows the multi-scale discriminator architecture in [Tang *et al.*, 2018; Zheng *et al.*, 2019]. During the training phase, each image is resized to 321 × 321, and then is random-cropped to 256 × 256. Randomly horizontally flipping is applied as simple data augmentation. We adopt Adam [Kingma and Ba, 2014] to optimize the whole IPM-Net with $\lambda_1 = 0.5$, $\lambda_2 = 0.999$ and set learning rate to 0.0001. We train our model for 1,000,000 iterations, and the batch size is set as 3.

Compared Approaches. To validate the effectiveness of the proposed method, we compare with the following baseline or alternative methods: (1) **CycleGAN.** We use CycleGAN [Zhu *et al.*, 2017] as the baseline model, which is widely applied for most style transfer tasks. We directly apply the open-source codes to the Makeup Transfer dataset. (2) **MUNIT** [Huang *et al.*, 2018] transforms the image into other domains by replacing the style features in the image, while the discriminative information in the original image is retained. Also, we train and test the open-source codes of MUNIT on the Makeup Transfer dataset. (3) **BeautyGAN** [Li *et al.*, 2018] also is a GAN-based model and performs fairly well on automatic makeup transfer. We perform qualitative and quantitative evaluations on the provided trained models.

4.2 Qualitative Results

Visualization. The qualitative results are shown in Fig. 5. The first row shows the makeup style reference images, while the first column shows the target face images. The result images are generated with the corresponding target image and the reference image. Our IPM-Net can effectively transfer the makeup styles provided by reference images for the non-makeup or light-makeup faces, while background and identity details are kept well. The makeup styles, *e.g.*, eyebrow, eye shadow, and the color of lipstick, are explicitly learned. We also test MUNIT [Huang *et al.*, 2018], BeautyGAN [Li *et al.*, 2018], and our model on the extra data collected from the Internet. The results are shown in Fig. 1. MUNIT fails to transfer the makeup style and leads to over-makeup and blurry images. We speculate that it is due to that MUNIT takes attention to the patterns of the entire images. MUNIT, therefore, is prone to transfer the general color mean and std to the target images. Since MUNIT has no local operations to pay attention to the small alteration, MUNIT also ignores some discriminative makeup styles, *e.g.*, the color of lipstick, and modifies the unrelated parts, *e.g.*, background. BeautyGAN generates two images with the original identity information, but the generated images contain limited makeup style translation. Besides, the backgrounds are also generally transformed, which leads to unrealistic artifacts. In contrast, the proposed IPM-Net leverages the image pre-processing methods to focus on the makeup area. The framework is further learned to disentangle the image into two codes. In the in-

Methods	Realism (FID) ↓	Diversity (LPIPS) ↑
Real	17.23	0.616
CycleGAN [Zhu *et al.*, 2017]	79.79	0.457
MUNIT [Huang *et al.*, 2018]	82.67	0.462
BeautyGAN [Li *et al.*, 2018]	56.57	0.459
Ours	**41.47**	**0.488**

Table 2: Comparison of FID (lower is better) and LPIPS (higher is better). We evaluate the realism and diversity of the makeup images and generated images on Makeup Transfer.

ference phase, we preserve the identity content code and only change the makeup-relevant part, *i.e.*, makeup style code, to generate the face with a specific makeup style. Since the identity content is kept, the discriminative identity information is well preserved. In this way, the background of the synthetic images also well remains. Besides, since the makeup style code of the proposed method focuses on the makeup part of the reference image, the detailed makeup patterns are also learned, including the color of lipstick and face foundation. As a result, the generated images of our method are more realistic and reflect the correct identity information and the discriminative makeup styles.

Controllable Makeup. In real-world scenarios, not all users need heavy makeup. It demands a controllable makeup inference, which motivates us to deploy a new reference makeup style code S_{new}. We set S_{new} as the weighted sum of the original makeup style code S_x and the reference makeup style code S_y, which could be formulated as:

$$S_{new} = \alpha S_x + (1 - \alpha) S_y, \tag{9}$$

where $\alpha \in [0, 1]$. We could control the value of α to manipulate the makeup transfer between light makeup and heavy makeup. The results are shown in Fig. 4. The generated results are sorted from left to right according to the makeup transfer level from light to heavy. It also verifies that the proposed model does not "memorize" the specific makeup style.

4.3 Quantitative Results

Accuracy. Except for the conventional qualitative results, we introduce the Fréchet Inception Distance (FID) [Heusel *et al.*, 2017] for image realism evaluation and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018] for generation diversity evaluation. We compare our method with the three competitive methods, *i.e.*, CycleGAN [Zhu *et al.*, 2017], MUNIT [Huang *et al.*, 2018], and BeautyGAN [Li *et al.*, 2018] under the same experimental setting. Specifically, we apply the automatic makeup method to generate makeup faces. The entire set of non-makeup images is transfer to the random makeup style. The makeup style is randomly selected from reference images in the makeup image set. To accurately report the results, we test each method for ten times and report average accuracy on FID (lower is better) and LPIPS (higher is better). The final results are shown in Table 2. Our method arrives 41.47 FID score, which is significantly lower than the second runner, *i.e.*, BeautyGAN [Li *et al.*, 2018], who is 56.57. The low FID score suggests our method is closer to the real images among all four methods. The results are also consistent with the former qualitative results. For generation diversity, our method also achieves a better LPIPS score, 0.488, than all other methods. The LPIPS of the proposed method is higher than the second runner, *i.e.*, MUNIT [Huang *et al.*, 2018] about 0.026. The improvement value is small, but it is relatively large. Comparing with the score between the generated image with the real images, which is $0.488/0.616 = 79.2\%$ and $0.462/0.616 = 75\%$, we yield about 4.2% improvement over the previous methods. It verifies that the proposed method could generate face with various markup styles. The model does not overfit or "memorize" some specific makeup style to cheat the discriminator.

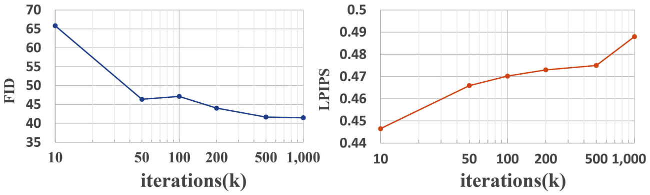


Figure 6: The FID and LPIPS curves with the increasing of different iteration. The results suggest that the proposed method converges stably and could gradually improve generation quality.

Convergence. Here we show the convergence of the proposed method. As shown in Fig. 6, the learning process of the proposed method is stable, and the IPM-Net converges well. It verifies the effectiveness of the proposed loss terms. When training with more iterations, the realism and diversity of the generated images are improved.

Speed. The image pre-processing for one input image takes 0.032 seconds. For the network inference, our model only needs 0.030 seconds to complete the automatic makeup for one target image.

5 Conclusion

In this paper, we propose a new model called Identity Preservation Makeup Net (IPM-Net) for the real-world automatic makeup problem. Our approach addresses the two real-world problemsin the existing automatic markup methods. It effectively transfers the makeup style of the reference image to the target image, while preserving the background as well as the critical patterns of the original identity. Controllable makeup transfer levels and fast processing make our model reaches one step closer to the real-world practice. Furthermore, both qualitative and quantitative experiments demonstrate that our method could achieve competitive results in real-world scenarios. In the future, we will further explore the image semantic understanding [Yan *et al.*, 2019; Luo *et al.*, 2005] to enhance the automatic face makeup.

Acknowledgments

This work is supported by National Nature Science Foundation of China (61931008, 61671196, 61701149,61801157, 61971268, 61901145, 61901150, 61972123), the National Natural Science Major Foundation of Research Instrumentation of PR China under Grants 61427808, Zhejiang Province Nature Science Foundation of China (LR17F030006, Q19F010030), 111 Project, No.D17019.

References

[Cao *et al.*, 2018] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018.

[Chang *et al.*, 2018] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *CVPR*, 2018.

[Chen *et al.*, 2019] 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*, 2019.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurlPS*, 2014.

[Guo and Sim, 2009] Dong Guo and Terence Sim. Digital face makeup by example. In *CVPR*, 2009.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurlPS*, 2017.

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

[Huang *et al.*, 2018] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015.

[Jiang *et al.*, 2019] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose-robust spatial-aware gan for customizable makeup transfer. *arXiv:1909.06956*, 2019.

[Jin *et al.*, 2019] Xin Jin, Rui Han, Ning Ning, Xiaodong Li, and Xiaokun Zhang. Facial makeup transfer combining illumination transfer. *IEEE Access*, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[Lai *et al.*, 2018] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *TPAMI*, 2018.

[Leyvand *et al.*, 2006] Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. Digital face beautification. In *ACM Siggraph*, 2006.

[Li *et al.*, 2015] Chen Li, Kun Zhou, and Stephen Lin. Simulating makeup through physics-based manipulation of intrinsic image layers. In *CVPR*, 2015.

[Li *et al.*, 2018] 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 *ACM MM*, 2018.

[Liu *et al.*, 2016] Si Liu, Xinyu Ou, Ruihe Qian, Wei Wang, and Xiaochun Cao. Makeup like a superstar: Deep localized makeup transfer network. *arXiv:1604.07102*, 2016.

[Luo *et al.*, 2005] Jiebo Luo, Andreas E Savakis, and Amit Singhal. A bayesian network-based framework for semantic image understanding. *Pattern recognition*, 38(6):919–934, 2005.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[Pu *et al.*, 2008] Yifei Pu, Weixing Wang, Jiliu Zhou, Yiyang Wang, and Huading Jia. Fractional differential approach to detecting textural features of digital image and its fractional differential filter implementation. *Science in China Series F: Information Sciences*, 51(9):1319–1339, 2008.

[Senft and Baym, 2015] Theresa M Senft and Nancy K Baym. What does the selfie say? investigating a global phenomenon. *International journal of communication*, 2015.

[Shashua and Riklin-Raviv, 2001] Amnon Shashua and Tammy Riklin-Raviv. The quotient image: Class-based re-rendering and recognition with varying illuminations. *TPAMI*, 2001.

[Sorokowski *et al.*, 2015] Piotr Sorokowski, Agnieszka Sorokowska, Anna Oleszkiewicz, Tomasz Frackowiak, Anna Huk, and Katarzyna Pisanski. Selfie posting behaviors are associated with narcissism among men. *Personality and Individual Differences*, 2015.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[Tang *et al.*, 2018] Wei Tang, Teng Li, Fudong Nian, and Meng Wang. Mscgan: Multi-scale conditional generative adversarial networks for person image generation. *arXiv:1810.08534*, 2018.

[Tong *et al.*, 2007] 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)*, 2007.

[Yan *et al.*, 2019] Chenggang Yan, Liang Li, Chunjie Zhang, Bingtao Liu, Yongdong Zhang, and Qionghai Dai. Cross-modality bridging and knowledge transferring for image understanding. *IEEE Transactions on Multimedia*, 21(10):2675–2685, 2019.

[Yu *et al.*, 2018] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[Zhang *et al.*, 2019] Honglun Zhang, Wenqing Chen, Hao He, and Yaohui Jin. Disentangled makeup transfer with generative adversarial network. *arXiv:1907.01144*, 2019.

[Zheng *et al.*, 2019] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.