The Visual Computer (2025) 41:173–189
https://doi.org/10.1007/s00371-024-03317-2

ORIGINAL ARTICLE

MakeupDiffuse: a double image-controlled diffusion model for exquisite makeup transfer

Xiongbo Lu[1] · Feng Liu[1] · Yi Rong[1] · Yaxiong Chen[1] · Shengwu Xiong[2,3,4]

Accepted: 10 February 2024 / Published online: 18 March 2024

Abstract

Makeup transfer is a challenging task, involving the transfer of a reference makeup style onto the source face while preserving the original appearance. Current GAN-based methods, representing makeup styles through reduced-dimensional matrices, often generate smooth high-frequency attributes and imprecise images. Additionally, these models are difficult to train and prone to model collapse problems. This paper introduces MakeupDiffuse, a diffusion-based model adapting a foundational diffusion model pre-trained on large-scale image datasets for makeup transfer. Fine-grained makeup transfer is achieved through a novel double image controller, controlling the identity process and adjusting the makeup style. To address the lack of paired data, we include another pre-trained makeup transfer network as a teacher module to supervise model training. Due to the flexible architecture, our model efficiently trains and generates faces with high perceptual quality and the expected makeup style. Unlike subjective evaluations based on user studies, we propose a new objective and comprehensive quantitative metric: Makeup Transfer Score, improving the current evaluation system. Extensive experiments demonstrate our approach's ability to generate natural makeup faces with exquisite details, achieving state-of-the-art performance.

Keywords Makeup transfer · Diffusion model · Controllable generation · Teacher module

1 Introduction

Makeup can enhance a person's appearance significantly, but it could be a time-consuming process to apply. Finding the right makeup products that work well for each individual will also take longer. Therefore, an automated system to transfer makeup styles from reference images onto our expected faces would be more practical.

Recent advances in makeup style transfer can be broadly categorized into two groups: conventional image processing techniques [1, 2] and methods based on deep neural network learning [3–15]. Traditional image processing methods typically involve decomposing an image into multiple layers, such as color and skin, and then utilizing the properties of physical reflection models to manipulate these layers for transferring information between them. Combining this approach with gradient operators enables the achievement of the final makeup transfer. Deep learning techniques, specifically GAN-based models [16], have been extensively employed for makeup transfer. These models usually conceptualize the makeup transfer and recovery processes as a closed-loop system, addressing the issue of limited paired data. Typically, deep learning techniques involve con-

Feng Liu's as Co-first author.

✉ Shengwu Xiong
xiongsw@whut.edu.cn

Xiongbo Lu
luxiongbo@whut.edu.cn

Feng Liu
liuf@whut.edu.cn

Yi Rong
yrong@whut.edu.cn

Yaxiong Chen
chenyaxiong@whut.edu.cn

1 The School of Computer and Artificial Intelligence, Wuhan University of Technology, 122 Luoshi Road, Wuhan 430070, Hubei, China

2 Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China

3 Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China

4 School of Information Engineering, Wuhan Huaxia Institute of Technology, Wuhan 430223, China

structing a deep neural network model that takes both a non-makeup face image and a reference makeup face image as input and utilizes this network to generate makeup-transferred results through direct training. Compared to traditional image processing methods, deep learning-based approaches are more resilient and can transfer makeup in complex scenarios with changing styles without requiring an accurate image decomposition. Traditional methods may occasionally fail to produce satisfactory outcomes, hence the current preference for deep learning methods.

GAN-based makeup transfer models primarily adopt the CycleGAN framework [17], trained on unpaired non-makeup and makeup images with additional supervision. This includes makeup loss to guide the reconstruction of specified makeup on the source face and identity loss to preserve the identity information of the source face. Despite their demonstrated exceptional makeup transfer ability, existing methods generally consider makeup styles as color distributions and often overlook details. Most GAN-based methods represent makeup styles using reduced-dimensional matrices, potentially resulting in smooth high-frequency attributes and imprecise generated images. Furthermore, GAN-based models are challenging to train and are frequently associated directly with model collapse problems.

Recently, the development of diffusion-based models has marked a breakthrough in image and text-to-image synthesis [18–22]. These methods typically involve two steps: diffusion with gradual noise addition in the forward direction and denoising in the subsequent term. The objective of these models usually revolves around learning the noise of the current diffusion step using a U-Net-like architecture [23]. Diffusion-based models have demonstrated an ability to capture semantic information [24] more easily and generate images with better sample quality than GANs. This superiority stems from their foundation in maximizing likelihood [25], which effectively avoids the mode collapse problem associated with GANs [16]. Such models have also achieved state-of-the-art performance in various common vision tasks, including super-resolution [26], image inpainting [27], sketch-based image translation [28], image-guided translation [29], predicting meningioma grade [30], etc. However, few works have concentrated on extending diffusion-based models for makeup transfer, a task that necessitates instance-level facial identity and makeup style controls rather than single-condition generation. This specificity underscores the need for our task to be more granular and detailed than a mere visual representation.

To address the aforementioned issues, we propose a novel diffusion-based [18] model named MakeupDiffuse for exquisite makeup transfer. Our model is constructed based on a foundational diffusion model pre-trained on large-scale image datasets. During training, we employ a Double Image Controller to preserve identity information and adjust the makeup style. Additionally, due to the lack of paired data, we have introduced a teacher module to supervise the model training. With a flexible design, our model can train efficiently, and, more importantly, it can generate high perceptual quality face images with the desired makeup style (Fig. 1).

Moreover, the existing evaluation of the makeup transfer task's effectiveness mainly falls into two types. The first is qualitative evaluation, where the makeup transfer performance is assessed manually. The second category is quantitative evaluation, akin to a user study, where several researchers are needed to distinguish the final generated facial makeup images. Both methods are inherently subjective. Hence, we propose a novel, objective, and comprehensive quantitative evaluation metric based on background consistency, global structure maintenance, and facial makeup style transfer, named Makeup Transfer Score (MT-Score).

By investigating the above ideas, the main contributions of our work can be summarized as follows:

- We propose a novel double image-controlled diffusion for exquisite makeup transfer, enabling fine detail control through a simple model. To address the challenge of missing paired data, we integrate a teacher network into our model to facilitate training and transfer.
- To the best of our knowledge, this is the first work that integrates the diffusion model with makeup transfer. Compared to the previous controllable diffusion model, this task requires ensuring content and style, significantly reducing randomness. Furthermore, our model generation is also faster and can be generated even in one step.
- This paper introduced a novel quantitative metric, MT-Score, designed to comprehensively measure the three crucial aspects of the makeup transfer task: background, face, and makeup. This index improves the evaluation system in the current research field to some extent.
- Our model can take full advantage of the pre-trained big model and teacher network, achieving state-of-the-art performance in makeup transfer.

The rest of the paper is organized as follows: We first briefly summarize Makeup Style Transfer, Controllable Diffusion Model, and existing Makeup Transfer Metrics that are related to our work in Sect. 2. Then, Sect. 3 describes the proposed MakeupDiffuse method and Makeup Transfer Score (MT-Score) evaluation metrics. Section 4 shows the experimental results and analysis. The conclusion of the paper is given in Sect. 5. Finally, we concisely analyzed and introduced the limitation of MakeupDiffuse and the following work plan in Sect. 6.

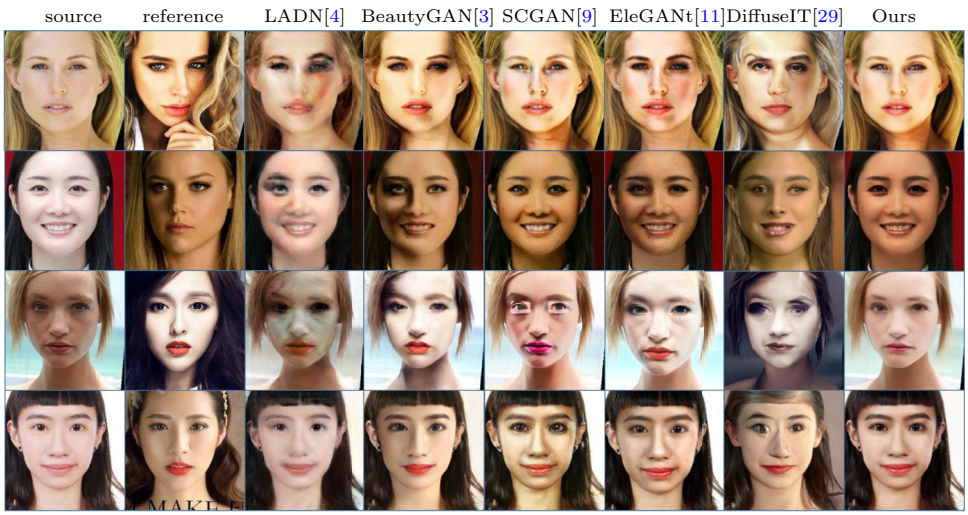

Fig. 1 Given some source-reference frontal image pairs (columns 1 and 2), our proposed method is qualitatively compared with the state-of-the-art methods. MakeupDiffuse could generate the most natural colors, shadows, and fewest artifacts, transferring results with the desired makeup style and high-quality details

2 Related work

2.1 Makeup style transfer

Facial makeup style transfer has become a topic of growing interest. This particular form of image-to-image translation has been the focus of research using GAN-based [16] methods to examine how makeup details can be extracted from reference images and transferred to target images. Typically, a separate encoder is employed to learn style representation and is concurrently optimized with a generator to apply makeup styles to target faces.

Inspired by Cycle-GAN [17], BeautyGAN [3] was proposed using an unsupervised manner for dealing with the lack of paired data, which incorporated a perceptual loss, a cycle-consistent loss, and a novel-designed makeup loss, resulting in instance-level makeup transfer and improved refinement of the makeup image. The same concept is also reflected in [13, 31], which emphasizes inter-domain differences and could transfer various local makeup styles such as eye shadow, lipstick, and foundation. Additionally, the end-to-end transfer of face makeup [14, 15] was found to be effective using a similar method. Jiang et al. [7] proposed a pose and expression robust spatially-aware GAN (PSGAN), which included an AMM module to extract the makeup matrix and control makeup shade. Although the results were not entirely satisfactory, this approach overcame the challenge of large pose and expression differences. Huang et al. [6] introduced IPM-NET, a face automatic makeup network based on the real world, which can automatically apply makeup while maintaining the character's identity and image background information, resulting in an authentic generated image. Nguyen et al. [8] addressed not only the color-changing of traditional makeup, such as eye-shadows and lipsticks, but also the migration of patterns, such as stickers, blushes, and jewelry, achieving the effect of in-the-wild makeup transfer. Deng et al. [9] proposed SCGAN, which divided the makeup transfer network into three components: target style-code encoding, face identity feature extraction, and makeup fusion. Sun et al. [10] focused on putting the makeup style to the corresponding semantic position of the target image and incorporated semantic correspondence learning to achieve makeup transfer and removal simultaneously.

Most of the above methods are based on GAN, which is difficult to train and often associated with model collapse problems, while we use another diffusion-based framework.

2.2 Controllable diffusion model

Jascha et al. [32] first propose the diffusion probabilistic model, which consists of two processes, a forward diffusion, and a reverse sampling process. The forward process, which sequentially samples the latent variables x_t for $t = 1, 2, 3, \ldots, T$, can be conceptualized as a Markov chain that gradually introduces noise to the data. Each step is a Gaussian transition: $q(x_t|x_{t-1}) \in \mathbb{N}(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I})$, where $\{\alpha_t\}_{t=0}^{T}$ is the hyperparameter that decreases over the step t. Utilizing the reparameterize trick, the step t diffusion variable x_t can be expressed as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}I\epsilon, \quad \epsilon \in \mathbb{N}(\mathbf{0}, \mathbf{I}) \tag{1}$$

where $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$.

Since then, some methods, for example, the probability density diffusion model (DDPM) [25], the implicit density diffusion model (DDIM) [24], and the score-based diffusion [33], have been proposed to improve the primitive model and speed up the model's training and sampling. Regarding image generation, compared to previous GAN-based methods, diffusion models can be trained and sampled directly at the RGB pixel level [25], the training is stable and the generated images are realistic. However, as the

image resolution increases, the generation efficiency of this method decreases significantly. To solve this problem, many researchers have carried out related research [20, 24, 34]. Among them, inspired by the latent image [35], the latent diffusion model (LDM) [18], including its upgraded version of stable diffusion (SD), was proposed and achieved state-of-the-art performance.

Given an image as a condition, many researchers focus on making the model controllable and producing the expected results [26, 27, 36, 37]. For instance, SRDiff [37] uses the diffusion model to realize the super-resolution of a single image. Adv-Diffusion [38] utilized strong inpainting capabilities of the diffusion model to generate imperceptible adversarial identity perturbations in the latent space. RePaint [27] employs a pretrained unconditional DDPM as the generative prior to implementing free-form inpainting. The diffusion model can accomplish the traditional image-to-image translation task well due to its friendly training and realistic generation. However, compared to these tasks, our study needs to meet two conditions simultaneously: facial identity and makeup style are guaranteed. The transferred results require more refinement and determinism without too much randomness. For example, although methods such as DiffusionClip [36] have proved in experiments that the makeup task can be achieved, its purpose is only to make up, not to specify what kind of makeup. The purpose of our makeup transfer task differs from those mentioned above and is more like a one-shot image-to-image translation. The closest research to our work is DiffuseIT [29], which implements image-guided style transfer and achieves state-of-the-art results. However, this task is domain level and the result of the transfer is more global. Our target is instance level, which requires more refined local details. Furthermore, our method is based on learning, while DiffuseIT can be regarded as optimization-based.

More recently, the quality of diffusion models for text-to-image synthesis and editing has reached previously unheard-of levels when big models are trained on pairings of text and images [18–20]. Control in these methods is usually achieved by encoding the prompt into latent vectors using pre-trained language models such as CLIP [39]. For instance, SDEdit [40] assumes that images in different domains can be mapped to the same diffusion step and edited from it. ControlNet [41] proposes a new conditional control framework and achieves conditioned control, edge maps, segmentation maps, keypoints, etc. Unlike the application scenarios of the above methods, makeup transfer can also be driven entirely by text. However, compared to having a makeup image as a reference, it is difficult and unnecessary to directly describe the makeup style in words and guide the source image for transfer.

2.3 Makeup transfer metrics

Like most generation problems, how to evaluate the quality of its generated results is still a problem worth exploring. Currently, the evaluation indicators of the effect of the makeup transfer task are mainly of two types. The first type is qualitative evaluation, that is, makeup transfer performance is judged manually. The second category is quantitative evaluation, similar to user study, meaning several researchers are selected to distinguish the final generated facial makeup images.

Qualitative evaluation The human eye is directly used to assess the results of the makeup transfer immediately [3, 4, 11]. This evaluation method is more subjective and focuses primarily on how well the overall makeup is natural and the local structure is managed. With the naked eye, details can be compared.

Quantitative evaluation User study refers to the selection of several researchers to evaluate the final produced facial makeup transferred images [7, 10]. This approach generally considered generated image quality, realism, and resemblance. This is the most popular evaluation technique used in the makeup transfer field.

The realization of the above two methods depends on the subjective judgment of people. Earlier studies [5] also proposed to compare the difference between the reference image and its cycle reconstructed one, such as structural similarity (SSIM), peak signal-to-noise ratio (PSNR), and mean-squared error (MSE). However, there is an extreme case of this method: when the network outputs the original input image without learning any makeup information, it instead achieves the best score.

3 Proposed method

3.1 Formulation

Let $\mathbf{X}$ and $\mathbf{Y}$ be the source domain and the reference domain, where $\{x^i\}_{i=1}^N$, $x^i \in \mathbf{X}$ and $\{y^j\}_{j=1}^M$, $y^j \in \mathbf{Y}$ donate the sample of non-makeup and makeup domain, respectively. The makeup transfer aims to learn a transfer function f: given x and y, $x_y = f(x, y)$, where x_y is the generated image with the face identity of x and the makeup style of y.

3.2 Network architecture

The architecture of double image-controlled diffusion (MakeupDiffuse) is so simple yet effective, as shown in Fig. 2, which mainly consists of four components: Source Face Encoder E_S, Reference Makeup Encoder E_R, Teacher Module T and Controllable Stable Diffusion D. Next, we will introduce these components in detail.

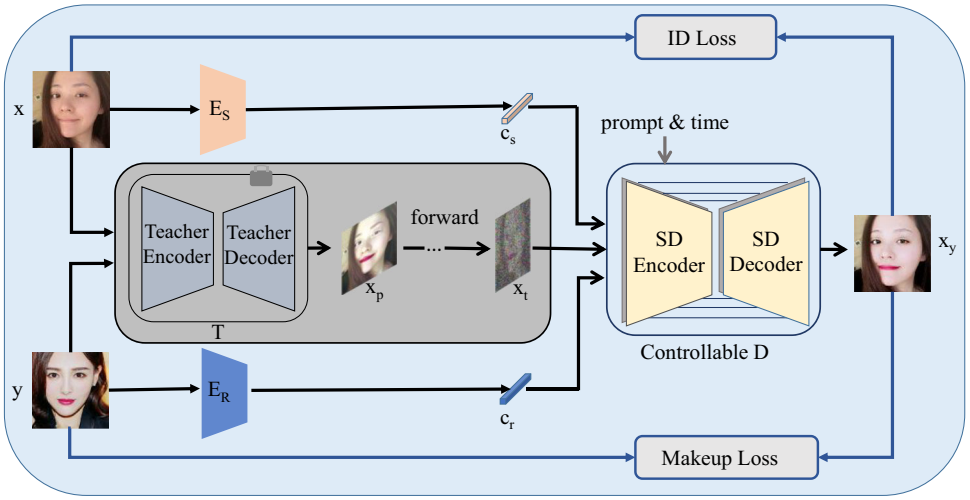

Fig. 2 The network framework of the MakeupDiffuse model. The source x and reference image y first pass through the fixed Teacher Module (T) to obtain a preliminary makeup transfer image x_p, then diffuse it and obtain latent x_t. Finally, we input the x_t and the content and style conditions (c_s, c_r) extracted by the trainable Source Face Encoder and Reference Makeup Encoder (E_S, E_R) into the Controllable Stable Diffusion Module (Controllable D) for a precise and controllable generation

3.2.1 Teacher module *T*

A significant problem in this makeup transfer is the lack of paired data, making it almost impossible to use fully supervised methods for learning effectively. The pre-trained teacher module T with fixed parameters can generate a preliminary transfer image x_p to solve the problem to a certain extent of no ground truth as follows:

$$x_p = T(x, y) \tag{2}$$

Although the generated x_p is imperfect, it can provide certain guidance for our model training. On the other hand, introducing this module can improve the generation speed of the next diffusion module. Our experimental design mainly chooses SCGAN [9] as our teacher module. But it should be noted that the teacher network is only used to quickly assist in generating x_p, and its choice is not unique from later experiments Sect. 4.4.

3.2.2 Feature encoder

The output of the teacher network may have problems such as information loss, artifacts, and low image quality. To make the final result not dominated by the teacher network, at the same time, to obtain additional original information supplementation, we increase the control of face and makeup.

In our method, E_S encodes the source non-makeup facial x into the facial content condition feature map c_s, which only contains the facial ID information of the image, while E_R extracts the makeup control condition c_r from the reference image y. These two encoders E_S and E_R have the same structure with [42], including three convolution blocks and six ResNet blocks. The kernel size of the first convolutional layer is set to 7, and the following two layers have a kernel size of 3. All of the strides in the convolutional layers are set to 1.

The above two conditional encoding processes can be formally described as follows:

$$c_s = E_S(x), \quad c_r = E_R(y) \tag{3}$$

3.2.3 Controllable stable diffusion module *D*

It is not easy to control the image operation directly in the original RGB space. At the same time, the final transferred image and the preliminary migration image have a certain degree of similar ID and makeup style characteristics. Therefore, we assume that x_p and x_y will be mapped to the same space, or even the same diffusion image, after t-step diffusion, which processing can be described as Eq. (4).

$$x_t = \sqrt{\bar{\alpha}_t} x_p + \sqrt{1 - \bar{\alpha}_t} \mathbf{I} \epsilon_t. \tag{4}$$

where ϵ_t is noise sampled from a standard Gaussian distribution, $\bar{\alpha}_t$ is the time-dependent diffusion coefficient, which detailed definition could refer to [25, 43].

Finally, facial content condition c_s and makeup style condition c_r jointly guide pre-trained Controllable Stable Diffusion module D [18] to complete the denoising of after t step diffused latent image x_t and generate makeup transferred image x_y, as expressed by:

$$\hat{\epsilon}_t = D(x_t, c_s, c_r, t), \tag{5}$$

$$x_y = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \mathbf{I} \hat{\epsilon}_t}{\sqrt{\bar{\alpha}_t}}. \tag{6}$$

where $\hat{\epsilon}_t$ is the output of the Controllable Stable Diffusion module. We hypothesize that it represents the noise in x_t obtained by diffusing x_y for t steps.

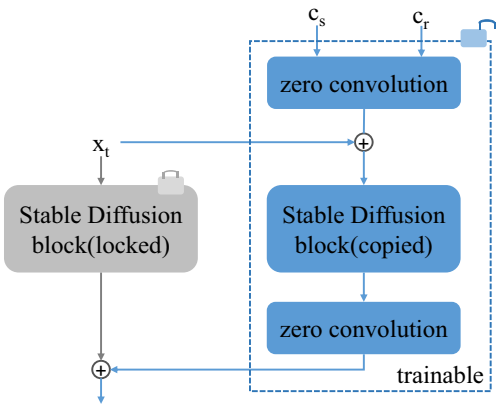

Fig. 3 Controllable stable diffusion block. The gray blocks represent stable diffusion modules that do not require training; in contrast, the blue module needs training

We adopted a network structure similar to [41], as shown in Fig. 3; each block contains a locked block and a replica trainable body of this block. Specifically, this paper only copies the image encoder and middle blocks of the stable diffusion model for training. Others such as the image decoder, prompt encoder, and time encoder remain fixed. Among them, the zero convolution module makes it possible to generate meaningful results even at the beginning of training. It makes the whole training process smooth until the convergence is completed. Different from previous work, our framework achieves double control of content and style. In addition, we do not expect the final result to have too much randomness. Hence, our MakeupDiffuse directly utilizes DDPM [25] to predict the noise and generate the image in one step without using the image generation process of DDIM [24, 36, 40].

3.3 Loss functions

Due to the solid generative ability of SD models trained on large-scale data and the help of the teacher module, our model differs from previous methods with many constraints and losses–instead, MakeupDiffuse only with Diffusion, ID, and Makeup loss.

$$L_{\text{total}} = \lambda_{\text{df}} L_{\text{DF}} + \lambda_{\text{id}} L_{\text{ID}} + \lambda_{\text{mk}} L_{\text{MK}} \tag{7}$$

where λ_{df}, λ_{id}, λ_{mk} control the relative importance of these three losses.

3.3.1 Diffusion loss

Since our model is based on the diffusion model, and the teacher network provides us with preliminary ground truth, we used the Diffusion loss in [25] to predict the noise in the diffused image x_t as defined as follows:

$$L_{\text{DF}} = ||\epsilon_t - \hat{\epsilon}_t||_2^2 \tag{8}$$

Optimizing this loss will produce images close to the output of the teacher network, which is slightly off for our purposes. Hence, we also need ID and makeup loss to guarantee the face and makeup so that the final result meets the makeup transfer task.

3.3.2 ID Loss

Inspired by the use of the reconstruction loss in the face identity disentanglement task, as discussed by Nitzan et al. [44], we adopt ID loss to constrain the final transferred result x_y to persist the ID information of the source image x.

$$L_{\text{ID}} = ||x_y - x||_2^2 \tag{9}$$

To avoid the generated image only learning facial identity information and ignoring makeup style, we only use this loss when the reference image y and source image x are the same ($y = x$) as expressed in Eq. (9).

3.3.3 Makeup loss

This loss is utilized to ensure that the generated image has the style of the reference makeup as defined as follows:

$$L_{\text{MK}} = \lambda_l L_{\text{lips}} + \lambda_e L_{\text{eye}} + \lambda_s L_{\text{skin}} \tag{10}$$

where L_{lips}, L_{eye}, and L_{skin} represent the lip, eye, and skin makeup transfer loss, respectively, defined in EleGANt [11]. λ_l, λ_e, λ_s are the weight values of these three losses, and following the settings in [11], their values are 10, 1, 0.5.

3.4 Makeup transfer score

The previous evaluation metrics are somewhat subjective, whether qualitative or quantitative analysis based on a user study. For this reason, this paper proposes the Makeup Transfer Score (MT-Score) evaluation method, which can be regarded as a valuable supplement to the existing research system. Before introducing the methodology, we need to answer a few questions. First of all, what is the purpose of makeup transfer? Analyzing the definition of makeup transfer, we expect the generated image to have the same background, structure, and facial identity as the original image and makeup as the reference image. What is the essence of makeup style that needs to be answered next? We think that the so-called makeup is a color distribution highly related to the structure of the face. For example, eye

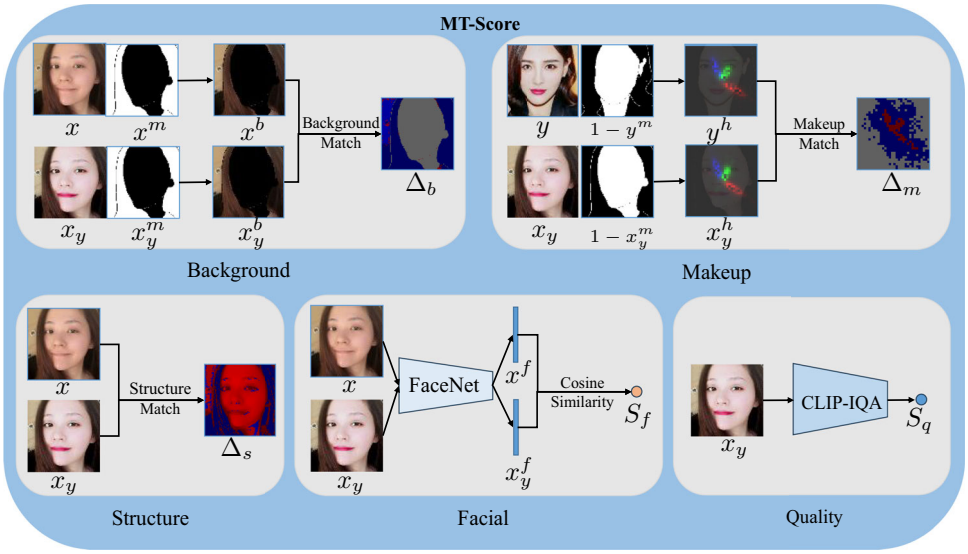

Fig. 4 Schematic diagram of the calculation of MT-Score. The description metrics are considered from five aspects of the generated image: background, makeup, structure, facial, and quality

makeup is the distribution of color around the eyes. Blush is a distribution of color around the cheekbones.

Perhaps it is slightly more complicated to describe the color distribution related to the structure directly. Fortunately, evaluating work does not necessarily mean finding a sufficient condition. To make the evaluation method more accessible, we can define the requirements that must be met for "excellent makeup transfer." That is, we only need to find the necessary conditions. As shown in Fig. 4, we will describe these necessary conditions separately below.

Background consistency In the makeup transfer task, the background of the source and generated image should be consistent. The ideal situation is precisely the same. Furthermore, since this item has genuinely ground truth, we expect the closer to the excellent score, the larger its gradient. Hence, we use the following formula to evaluate the background.

$$\Delta_b = \mathrm{MSE}\left(x \odot x^m, x_y \odot x_y^m\right) \tag{11}$$

$$S_b = 1 - 2 * \sigma\left(\log\left(s * \Delta_{\mathrm{back}}\right)\right) \tag{12}$$

where x^m and x_y^m represent the background mask obtained by x and x_y through the semantic segmentation network, respectively; s is a scaling factor less than 1, here we specify it as 0.5; $\odot$ represents element-wise multiplication; $\mathrm{MSE}(\cdot,\cdot)$ represents the MSE function; $\sigma(\cdot)$ represents the sigmoid function.

Makeup style transferring As previously described, we only evaluate the prerequisites for makeup. That is, whether the facial color distribution of the generated image is consistent with that of the reference image. Here, we employ the distribution histogram of the face color to express the makeup style without considering its relationship with the face structure. Evaluation can be formalized as follows:

$$\Delta_m = L_{\mathrm{hist}}\left(h\left(y \odot \left(1 - y^m\right)\right), h\left(x_y \odot \left(1 - x_y^m\right)\right)\right) \tag{13}$$

$$S_m = 1 - \Delta_{\mathrm{makeup}} \tag{14}$$

where y^m represents the background mask obtained by reference image y through the semantic segmentation network; $h(\cdot)$ can calculate the histogram of the input image; $L_{\mathrm{hist}}(\cdot,\cdot)$ is the histogram loss, which detailed definition could refer to [45].

Structure maintenance The generated image's global structure needs to be consistent with the source image. We utilize SSIM to measure this consistency. At the same time, avoiding generation without change is best. We propose the basis of the following metric on its significance will decrease when SSIM reaches a certain threshold.

$$\Delta_s = 1 - \mathrm{SSIM}\left(x, x_y\right) \tag{15}$$

$$S_s = \cos\left(\frac{\pi}{2}\Delta_{\mathrm{structure}}\right) \tag{16}$$

where $\mathrm{SSIM}(\cdot,\cdot)$ represents the SSIM function.

Facial persistence Like other face editing algorithms [12, 46], facial identity information should be preserved after makeup transfer. The measure of identity preservation is implemented by the cosine similarity of identity features, which are extracted by the pre-trained state-of-the-art face recognition model FaceNet [47]. It can be expressed formally as:

$$S_f = \frac{x^f \cdot x_y^f}{||x^f||_2 \cdot ||x_y^f||_2} \tag{17}$$

$$x^f = F(x), \quad x_y^f = F(x_y) \tag{18}$$

where $F(\cdot)$ represents the output of the trained FaceNet, which is the feature vector.

Quality goodness A good image after makeup transfer should satisfy a high quality overall, which should meet the requirements of both look and feel. In CLIP-IQA [48], the authors introduce human perception into the no-reference image evaluation system. Specifically, they input images described as good or bad to the evaluation model for training to make the model have good aesthetic perception ability. Finally, users can use the trained model to evaluate and score the input image. A higher score means higher quality and better feeling.

In our approach, we use CLIP-IQA for reference-free evaluation of image quality. It can be formally described as:

$$S_q = M(x_y) \tag{19}$$

where $M(\cdot)$ is the output of the CLIP-IQA model, which is the CLIP-IQA score.

Generally, the above five sub-evaluation indicators can be divided into two levels of pixel and perception. Background Consistency, Makeup Style Transferring, and Makeup Style Transferring are evaluated at the pixel level. For the pixel-level score, we have the following definitions:

$$S_{\text{pixel}} = \frac{\lambda_b S_b + \lambda_m S_m + \lambda_s S_s}{\lambda_b + \lambda_m + \lambda_s} \tag{20}$$

where λ_b, λ_s, and λ_m control the relative importance of these three local pixel-level sub-metrics, we set them to 2, 3, 1, respectively.

Facial Persistence and Quality goodness belong to a higher level of perception evaluation. For the perceptual score, we have the following definition:

$$S_{\text{perception}} = \frac{\lambda_f S_f + \lambda_p S_p}{\lambda_f + \lambda_p} \tag{21}$$

where λ_f and λ_p control the relative importance of these two global perception-level sub-metrics S_f and S_p, and we set them to 2 and 1, respectively.

Based on satisfying the primary conditions, we believe the makeup transfer task pays more attention to the evaluation at the perceptual level. In the end, our MT-Score can be expressed as:

$$S = \sqrt{S_{\text{pixel}}} \cdot S_{\text{perception}} \tag{22}$$

4 Experiments

4.1 Experiment setup

4.1.1 Implementation details

In our experiments, Adam is used to optimize the MakeupDiffuse model for 20 epochs, and the learning rate is initialized to 1e−5. The hyperparameters in Eq. (7) are fixed at $\lambda_{\text{id}} = 1$, $\lambda_{\text{mk}} = 10$, $\lambda_{\text{df}} = 1$, diffuse steps are set to $t \in (100, 1000)$ and $t \in (100, 700)$ in the training and testing phase, respectively. Each test sample was taken three times for quantitative comparison, and the average value was recorded. All images are resized to 256×256 before training, and the prompt is set to "makeup transfer" to ensure that its variation does not impact the experiment and opens up the possibility for future text-driven makeup transfer scenarios. We implement our network using PyTorch and train it on CPU: E5-2609V4*2 @1.70 GHz; GPU: Nvidia GeForce RTX 3090 GPU*1; RAM: 128 GB; Ubuntu 20.04. For more details,please refer to our source code.[1]

4.1.2 Datasets

Consistent with the previous approach, the MT (Makeup Transfer) dataset [3], which contains 1115 non-makeup images and 2719 makeup images, including variations in poses, races, etc., is used to train our model and follow the strategy of [9] to split the train/test set. Moreover, to further validate the effectiveness of MakeupDiffuse for handling source and reference images with large poses and expressions, we incorporate the Makeup-Wild dataset [7] as an additional test set. This dataset contains facial images with various poses and expressions as well as complex backgrounds to test methods in the real-world environment. It is important to note that our network is exclusively trained using the training part of the MT dataset to ensure a fair comparison.

4.2 Comparisons with state of the arts

We conduct comparisons of MakeupDiffuse with some state-of-the-art makeup transfer methods: LADN [4], BeautyGAN [3], SCGAN [9], EleGANt [11]. Furthermore, we also compare our method with the diffusion-based style transfer model DiffuseIT [29]. For a fair comparison, the models of all these methods come from their official implementations, and we only test on this basis.

[1] https://github.com/jiean001/MakeupDiffuse.

4.2.1 Qualitative comparison

As demonstrated in Fig. 1, we show qualitative comparison results with baselines of frontal facial makeup transfer cases. We observe that although LADN [4] maintains certain face information and transfers part of the makeup, it produces too many artifacts, resulting in low-quality results. For example, the cheeks in the first image and the eyes in the second image have noticeable artifacts; the third image seems to transfer the color of the background sea to the face; the fourth image makeup transfer is not apparent; and the overall comparison is blurred. BeautyGAN [3] could produce images with makeup on; however, the result may have partial artifacts, obviously abnormal shadows, etc. For instance, the first image seems to transfer the shadow of the reference image along with the makeup; the second image has an abnormal shadow on the left side of the face, and its background has also changed from red to dark red. Due to its design dependence on the mask, the makeup transfer of SCGAN [9] in some details is not clear enough, e.g., the eye makeup of the third image rectangle, etc. DiffuseIT [29] may be better at dealing with domain-level style transfer tasks, which leads to its insufficiency in dealing with the need to focus more on local instance-level makeup transfer tasks. It generates results that look like they keep the face and transfer the makeup, but the result seems unreal, with obviously machine-generated artifacts, like a robot face or a human face in a game. Compared to the above methods, EleGANt [11] could produce relatively realistic and desired images. However, this method is implemented step by step, which needs to crop the face, perform makeup transfer, and finally utilize post-processing to obtain the final image. This operation tends to change the structure of the source image. The third and fourth images, such as image size changes, can prove this point. The generation of structural changes is inconsistent with the goal of style transfer. In addition, it also has a few artifacts, e.g., the right jaw in the first picture and the left side of the face in the second picture.

When compared to the above state-of-the-art methods, our method produces high-quality images with the most exact makeup styles, regardless of whether the comparison is in eye shadows, lipsticks, or foundations. In the third row, for example, only our result transmits the reference image's makeup style. The results suggest that MakeupDiffuse preserves other makeup-insignificant components such as hair, clothes, and backdrop as the original non-makeup images.

Figure 5 shows some non-frontal makeup transfer results in MT test datasets. The experimental results show that our method has achieved the most natural makeup transfer, miniature artifacts, and solid background and facial persistence, whether on the left face, right face, decorative glasses, or artificial occlusion. Combining the qualitative evaluation of the front and non-front makeup transfer, our method's effectiveness, generalization, and robustness are demonstrated.

We also compare results on the Makeup-Wild test set, where both source and reference images feature larger poses and expressions, as illustrated in Fig. 6. Current methods often lack a clear mechanism to guide pixel-level transfer direction or rely on face parsing methods and also overfit on frontal images. Consequently, in a real-world environment, makeup may be applied in the wrong region of the face. For instance, in the third row of Fig. 6, the lip gloss from an open-mouthed face is transferred to the skin. In the first row, other methods struggle to transfer makeup onto side faces. However, our proposed method, leveraging the robust prior ability of the pre-trained diffusion model, accurately assigns makeup to each pixel, resulting in visually superior outcomes.

4.2.2 Quantitative comparison

Following previous works, we conduct a user study to quantitatively evaluate the makeup style transfer precision and face identity persistence. We randomly selected 30 generated images from the test split of the MT dataset. Table 1 shows the results of 50 people to evaluate, where "ID", "Style", "Detail", and "overall" donate the four aspects: ID persistence", "makeup style transfer", "detail processing", and "overall performance". Our MakeupDiffuse method outperforms the others in "ID", "Details", and "Overall" respects, especially in exquisite detail generation.

In addition, we also quantitatively compare these methods using MT-Score. We randomly select 100 non-makeup images and three makeup reference images of each non-makeup face for comparison. The experimental results are shown in Table 2. This table shows that our proposed method achieves state-of-the-art performance in the two sub-items, e.g., "Background Consistency" 95.58, and "Facial persistence" 90.34, respectively. In terms of "Face Preservation", only our MakeupDiffuse is greater than 90, and it is nearly 4.8 points higher than the second place. In several other aspects, our method comes second, but the gap with the first place is not apparent. Especially in terms of "Structure Maintenance" and "Quality Goodness", the gap is less than 0.3. And even in "Makeup Style Transferring", which has the most significant gap, the gap is only less than 1.3. The results of the one pixel-level terms prove that our method can maintain the original background better than other methods. The evaluation indicators of the two perceptual-level are better than other methods, which proves that our method complies with the requirements of face preservation and the final result makes people feel good.

In the end, the MT-Score of our MakeupDiffuse method is 63.32, which is significantly better than other methods, and 1.56 higher than the second-place EleGANt method.

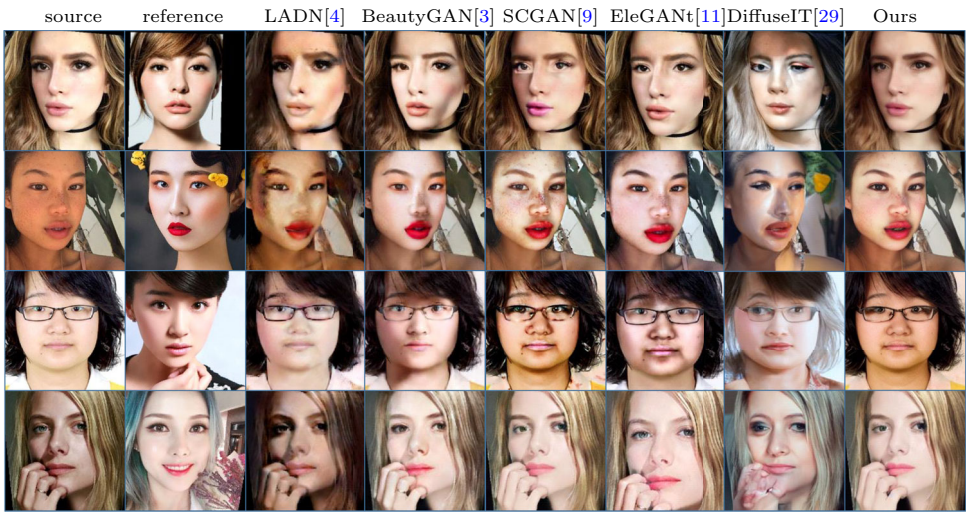

Fig. 5 Qualitative comparisons with some state-of-the-art approaches in the non-frontal cases. MakeupDiffuse is more robust and could generate the desired makeup transfer results regardless of side faces, decorations, or occlusions

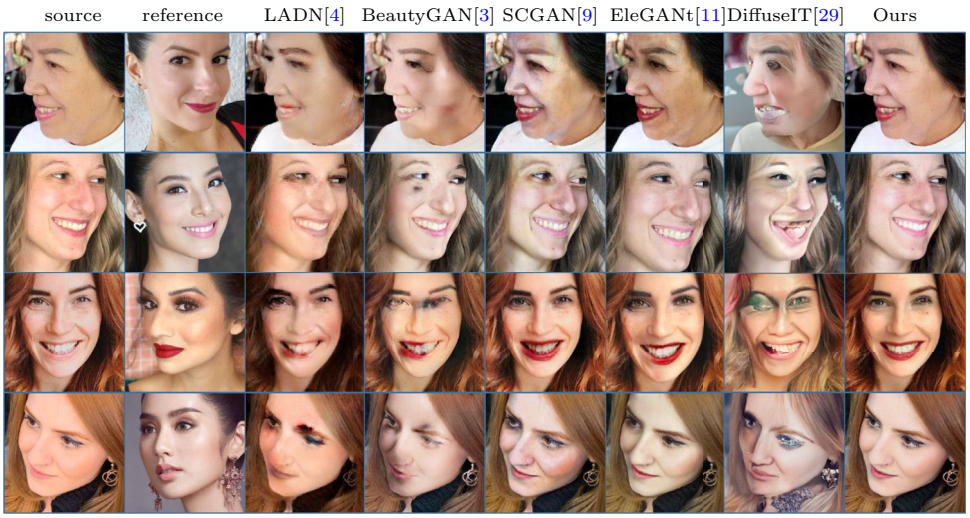

Fig. 6 The qualitative comparison was conducted on an additional Makeup-Wild test set, revealing that the performance of MakeupDiffuse is even more impressive when handling source and reference images with larger poses and expressions

Table 1 User study results(ratio (%) selected as the best)

	LADN [4]	BeautyGAN [3]	SCGAN [9]	EleGANt [11]	DiffuseIT [29]	Ours
ID	2.20	21.87	22.40	24.93	1.53	**27.07**
Style	2.07	21.40	24.13	**26.47**	0.73	25.20
Detail	1.27	22.53	23.40	24.73	0.54	**27.53**
Overall	1.73	22.00	23.33	25.20	0.94	**26.80**

The bold items indicate the best results achieved under the corresponding evaluation indicators

Table 2 MT-Score results of our MakeupDiffuse and existing state-of-the-art methods

	LADN [4]	BeautyGAN [3]	SCGAN [9]	EleGANt [11]	DiffuseIT [29]	Ours
S_b	92.62	94.31	94.41	91.53	82.71	**95.58**
S_m	64.40	76.68	78.60	**82.97**	68.39	81.69
S_s	92.22	**97.74**	97.27	92.41	71.04	97.53
S_f	72.66	77.48	82.35	86.56	53.83	**90.34**
S_q	31.83	49.43	49.46	**55.82**	38.82	55.54
MT-Score	40.24	54.53	56.36	61.76	37.60	**63.32**

The bold items indicate the best results achieved under the corresponding evaluation indicators

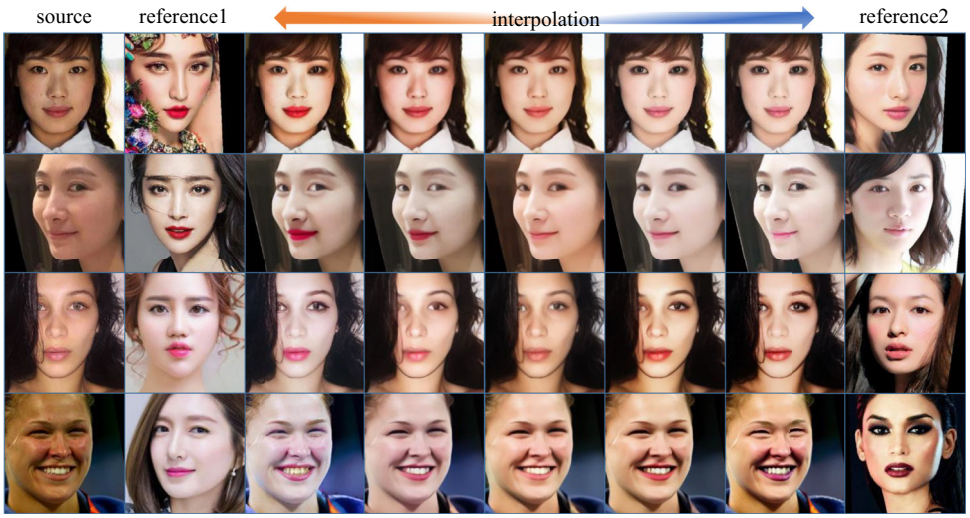

Fig. 7 Interpolated makeup transfer. Our proposed MakeupDiffuse could achieve smooth makeup interpolation transitions no matter in the case of a normal front face, partial face, occlusion, and obviously makeup style gap between reference and source face

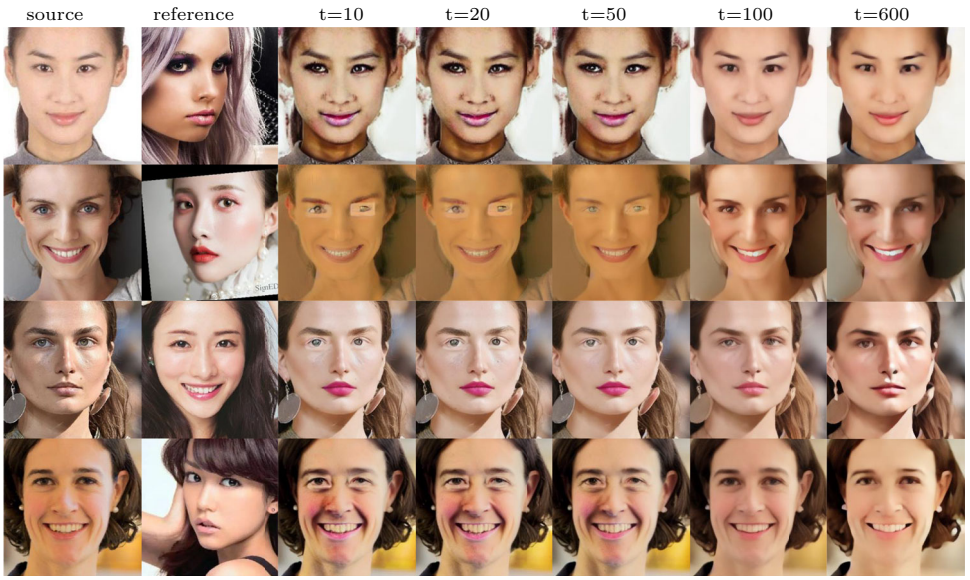

Fig. 8 In the training phase, the influence of the selection of diffusion step t on the experimental results. Too small step t will lead to unnatural generated results, we set $t \in (100, 1000)$ in this phase

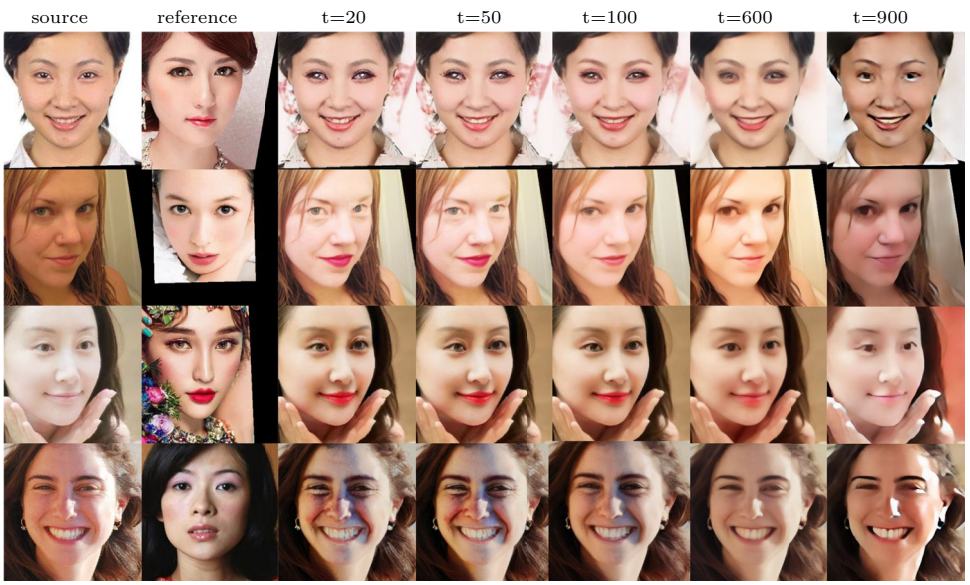

Fig. 9 Different diffusion step t choices in the generation stage influence the experimental results. Overall, the experimental results show a trend of suitable first and then bad; we set $t \in (100, 700)$ in our transfer phase

By comparing the evaluation of MT-Score in Table 2 to the assessment of humans in Table 1, it is found that the two are compatible. This proves the feasibility of our proposed MT-Score evaluation metrics.

4.3 Extended experiments

4.3.1 Controllable makeup transfer

Since we use an extracted makeup style vector c_r and preliminary transfer image x_p to control the makeup style result of the transferred image, it can easily control fusing the styles from multiple reference images with linear interpolation. This interpolation process can be described as followings:

$$x_p = \alpha x_p^{r_1} + (1-\alpha)x_p^{r_2}, \tag{23}$$

$$c_r = \alpha c_r^{r_1} + (1-\alpha)c_r^{r_2}. \tag{24}$$

where $\alpha \in (0, 1)$ is the interpolation coefficient, r_1 and r_2 represent two different referent images.

We randomly select two makeup styles, perform linear interpolation on their style features and preliminary transfer images, and then use the interpolated features to control the target makeup. The experimental results are shown in Fig. 7, demonstrating that our method can achieve smooth makeup transitions flexibly and controllably no matter in the case of a normal front face, partial face, occlusion, and obviously makeup style gap between reference and source face.

4.3.2 The selection of diffusion step

We tested the effect of diffusion step t on makeup transfer in two phases, as shown in Figs. 8 and 9. Overall, combined with the MT-Score evaluation in Tables 3 and 4, the experimental results show a trend of suitable first and then bad. The smaller t is, the richer makeup colors can be displayed, but it will affect the ID information and generate some artifacts. With the increase of t, it becomes more and more difficult for the face to learn the target information, and the generated image does not look realistic enough. In theory, there is an optimal training and sampling step t. But to achieve simplicity, result flexibility, and generalization of the method, this paper limits the range of it to (100, 1000) in training and (100, 700) in the transfer phase.

4.4 Ablation studies

4.4.1 Effects of teacher module

To further evaluate the teacher module, in addition to choosing the SCGAN network [9] as the teacher module, we also adopt the pseudo-ground-truth method proposed in EleGANt [11] as the teacher module for experiments. The experimental results are shown in Fig. 10 and Table 5. Compared to the preliminary makeup transfer image generated by the teacher module (e.g., the third and fifth columns), our transfer results (e.g., the fourth and sixth columns) are more natural and aligned with expectations. For instance, SCGAN generated rectangular eye makeup, more artifacts, less accurate lip makeup, etc. With our method, the generated eyes are more natural and the makeup color is more in line with the requirements. The result of PGT obviously feels like the face is masked and the part of the face is obviously in sharp contrast to the background. It does some color migration, but it does not look natural overall. In contrast, our results clearly overcome the above defects. Quantitative analysis also proves our analysis; as can be seen in Table 5, our final MT-Score is six to seven points higher than the original makeup transfer model. When the teacher network has a certain enough generative ability, the results generated by our MakeupDiffuse are hardly affected by the choice of teacher module; even if the teacher module is different, our model can still produce similar results.

Furthermore, to demonstrate the necessity of the teacher network, we also tested the model without this module and used the source image as a surrogate for the preliminary makeup transfer image. Its qualitative results are shown in the last column of Fig. 10. We observe that although the background and face can be generated with a relatively correct structure and has a little makeup when this module is removed, its overall makeup information is lacking. The quantitative evaluation of the last column in Table 5 also con-

Table 3 MT-Score results of the different train diffusion step t

	$t = 10$	$t = 20$	$t = 50$	$t = 100$	$t = 600$
S_b	94.10	94.53	95.19	95.49	94.97
S_m	72.04	74.16	78.82	81.70	80.89
S_s	91.66	92.91	94.11	97.73	94.80
S_f	80.43	80.82	85.31	89.26	87.68
S_q	45.09	47.82	52.62	55.42	55.21
MT-Score	51.71	53.93	59.19	62.92	61.91

Table 4 MT-Score results of the different test diffusion step t

	$t = 20$	$t = 50$	$t = 100$	$t = 600$	$t = 900$
S_b	94.20	95.37	95.55	95.51	91.50
S_m	72.76	77.42	81.69	79.42	64.16
S_s	91.42	96.24	97.50	97.20	93.50
S_f	82.47	83.41	89.97	88.21	81.41
S_q	52.07	53.13	56.05	55.65	40.45
MT-Score	56.67	58.81	63.52	62.30	47.83

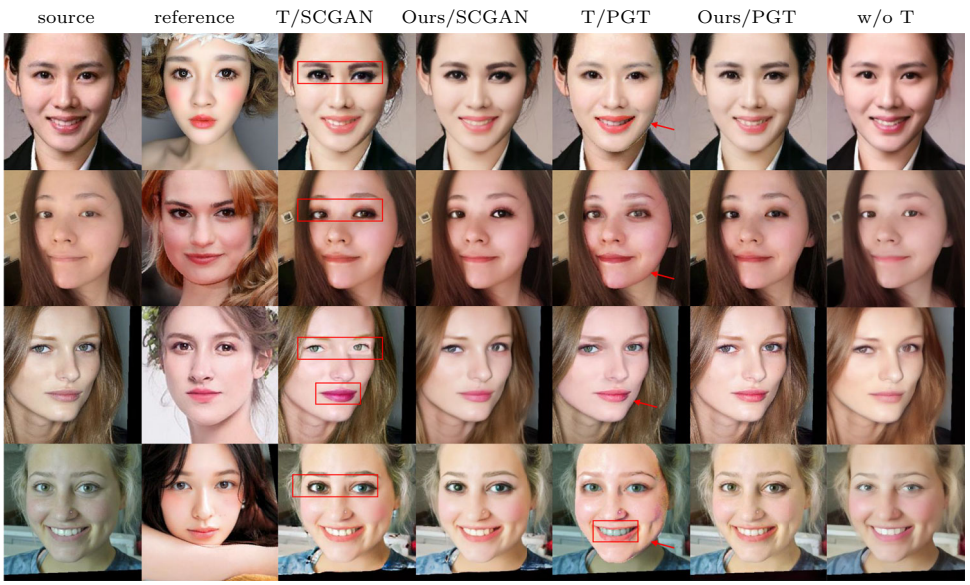

Fig. 10 The influence of different teacher modules and ablation study by removing this module on the experimental results. Compared with the artifacts of the teacher network, inaccurate colors, fine details, and masked faces, our model results are more accurate and natural. After removing the teacher network, although the background and face can be generated with a relatively correct structure and has a little makeup, overall makeup information is lacking

Table 5 MT-Score of the different teacher modules and ablation study by removing this module

	T/SCGAN	Ours/SCGAN	T/PGT	Ours/PGT	w/o T
S_b	94.38	95.57	99.32	95.59	95.99
S_m	78.40	81.53	85.95	81.73	63.23
S_s	97.20	97.48	99.41	98.51	97.96
S_f	82.91	90.25	83.21	91.52	90.63
S_q	49.18	55.76	44.59	54.93	51.86
MT-Score	56.32	63.40	55.31	63.38	57.92

Fig. 11 Ablation study by removing control module and loss one by one, where c_s, c_r, and c represent the face control module, the makeup control module, and the simultaneous face and makeup control module; L_{ID}, L_{MK}, and L represent the ID loss, the makeup loss, and the simultaneous ID and makeup loss. Removing different components will affect the final build, possibly with some unwanted consequences; for example, the facial is not persistent enough (more obvious ones are columns 3, 6, 7), or the makeup is missing (columns 3, 5) or mismatched (columns 2, 6, 7)

Table 6 MT-Score of the ablation study

	w/o c_s	w/o c_r	w/o c	w/o L_{ID}	w/o L_{MK}	w/o L	w/o $c\&L$	Ours
S_b	94.06	94.36	94.06	94.28	94.36	94.26	94.05	95.49
S_m	80.17	77.72	76.68	79.87	76.76	76.64	76.52	81.23
S_s	95.68	95.97	95.70	95.83	95.94	95.82	95.68	97.72
S_f	77.71	83.80	77.12	80.81	83.18	80.58	76.55	89.78
S_q	54.34	54.37	53.47	54.28	54.15	53.34	53.01	55.88
MT-Score	58.08	59.62	56.78	58.99	59.13	57.79	56.29	63.28

Removing the face control condition c_s or ID Loss L_{ID} will affect the "Facial Persistence" of the final generated result. Removing the makeup control condition c_r or Makeup loss L_{MK} will affect the "Makeup Style Transferring." As the two control modules c, the two losses L, and both the control module and the loss $c\&L$, are removed one by one, the resulting MT-Score gets lower and lower

Fig. 12 Ablation study by removing the control modules for testing. This blurred image and unrefined makeup show from the side that when the image is generated, the face and makeup information is injected through the control module. It confirms the necessity of our control module, and the teacher network cannot dominate the generation

firms this. The score of the "makeup style transfer" subitem is only 63.23, and the result lags far behind.

In summary, it can be concluded that our results depend on the teaching network. Still, the teacher network cannot dominate the entire generation, and our results are significantly better than the initial output of the teacher network.

4.4.2 Effects of controller module

Since the generated preliminary transfer image is imperfect and may lack information, we could not allow the teacher module to dominate the result. In addition, we increase the control of face and makeup to obtain additional information supplementation. To verify our idea, we performed the following ablation experiments: (1) removing the ID control module c_s while keeping losses; (2) removing the makeup control module c_r while keeping the losses; (3) removing both ID and makeup control module c while keeping the losses.

The qualitative and quantitative evaluation of the experiment is shown in Fig. 11 and Table 6, respectively. No matter which module is removed, the experimental results will be reduced, but the degree of reduction is different. The absence of the ID module significantly reduced the "Facial Persistence" sub-metrics. Its value suddenly dropped from 89.78 to 77.71. The lack of the Makeup module significantly lowered the "Makeup Style Transferring" evaluation of the final result, and its value dropped by 3.51. The absence of a control module means that the final generation results depend entirely on the teacher network. The evaluation of each subitem has dropped significantly, and the final MT-Score is only 59.62. It should be noted that this score is higher than the scoring output by the teacher network itself because the powerful diffusion model will significantly improve the "Quality Goodness" sub-metrics score of the generated results.

To prove the necessity of the control module on the other side, we removed E_S and E_R for testing. The experimental results are shown in Fig. 12, which demonstrates the comparison of our results with the removal of the control module. It can be seen that the generated images are blurred, the background and makeup are not clear, and the mosaic effect is noticeable. This suggests that the teacher module cannot fully provide the information we want. It is also confirmed that the Source Face Control and Reference Makeup Control modules can effectively constrain the generated images.

4.4.3 Effects of loss functions

Compared with the original diffusion model, we introduce ID and Makeup loss for the makeup transfer task. We performed the following ablation experiments to demonstrate their effectiveness: (1) removing the ID Loss L_{ID} while keeping the double controls; (2) removing the makeup control Loss L_{MK} while keeping the double controls; (3) removing both ID and makeup losses L while keeping the double controls. Moreover, we add (4) we are fine-tuning the corresponding Stable Diffusion blocks without condition injections to prove our modules and losses.

Observing the quantitative evaluation Table 6 and qualitative analysis Fig. 11, it is found that after removing the ID loss, its "Facial Persistence" sub-metrics dropped by nearly nine points. When the makeup loss is removed, its "Makeup Style Transferring" sub-item score drops from 81.23 to 76.76. When both losses are deducted, the final result falls by 5.49. Only fine-tuning the corresponding Stable Diffusion blocks achieves the worst performance, which is also lower than the output of the teacher network itself.

5 Conclusion

In this paper, we employ the diffusion model for makeup transfer for the first time and demonstrate it can achieve both content and style control. We propose the MakeupDiffuse model with two pre-trained modules and two dependent conditions control, e.g., Source Face and Reference Makeup control. The Teacher Module could generate a preliminary makeup transfer image, which eases the lack of paired data

and accelerates the generation processing. The Controllable Stable Diffusion module can control the transferred image's ID and makeup style. With the help of stable diffusion power generation capabilities, our transferred results are more natural, exquisite, as expected, and have almost no artifacts. There is a difference between the previous method of the step-by-step denoising generation. After predicting the noise of our model, we can directly generate the final image in one step. Moreover, we propose a novel objective and comprehensive Makeup Transfer Score quantitative metrics to improve the existing evaluation system.

6 Limitation and future work

We believe that the double image-controlled diffusion framework would be helpful for other controllable generations and tasks. Besides, similar to introducing the Teacher and Controllable Stable Diffusion Modules, many trained models will also be applied to future model design in a plug-and-play manner. While our model has demonstrated successful makeup transfer, it currently encounters challenges in handling very large poses, exaggerated expressions, and partial makeup transfer in real makeup trial scenes. Consequently, our next research endeavor aims to incorporate text control to enhance the applicability of the makeup transfer algorithm in complex real-world scenarios . In addition, inspired by some other work related to face synthesis [49, 50], we will also strive to achieve makeup transfer conditioned on sketches or heterogeneous face images.

Funding This work was in part supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160604) and NSFC (Grant No. 62176194), and the Key Research and Development Program of Hubei Province (Grant No. 2023BAB083), the Project of Sanya Yazhou Bay Science and Technology City (Grant No. SCKJ-JYRC-2022-76, SKJC-2022-PTDX-031), and the Project of Sanya Science and Education Innovation Park of Wuhan University of Technology (Grant No. 2021KF0031).

Data availability Data sharing is not applicable to this article as no datasets were generated during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Guo, D., Sim, T.: Digital face makeup by example. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 73–79 (2009)
2. Li, C., Zhou, K., Lin, S.: Simulating makeup through physics-based manipulation of intrinsic image layers. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4621–4629 (2015)
3. 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 ACM Multimedia Conference on Multimedia Conference (ACM MM), pp. 645–653 (2018)
4. 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 International Conference on Computer Vision (ICCV), pp. 10480–10489 (2019)
5. Zhang, H., Chen, W., He, H., Jin, Y.: Disentangled Makeup Transfer with Generative Adversarial Network (2019)
6. Huang, Z., Zheng, Z., Yan, C., Xie, H., Sun, Y., Wang, J., Zhang, J.: Real-world automatic makeup via identity preservation makeup net. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)
7. 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 International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5193–5201 (2020)
8. Nguyen, T., Tran, A.T., Hoai, M.: Lipstick ain't enough: beyond color matching for in-the-wild makeup transfer. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13305–13314 (2021)
9. Deng, H., Han, C., Cai, H., Han, G., He, S.: Spatially-invariant style-codes controlled makeup transfer. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6549–6557 (2021)
10. Sun, Z., Chen, Y., Xiong, S.: SSAT: a symmetric semantic-aware transformer network for makeup transfer and removal. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 2325–2334 (2022)
11. Yang, C., He, W., Xu, Y., Gao, Y.: Elegant: exquisite and locally editable GAN for makeup transfer. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 737–754 (2022)
12. Tiwari, H., Subramanian, V.K., Chen, Y.-S.: Real-time self-supervised achromatic face colorization. Vis. Comput. 1–16 (2022)
13. Organisciak, D., Ho, E.S., Shum, H.P.: Makeup style transfer on low-quality images with weighted multi-scale attention. In: Proceedings of the International Conference on Pattern Recognition (ICPR), pp. 6011–6018 (2021)
14. Lyu, Y., Dong, J., Peng, B., Wang, W., Tan, T.: SOGAN: 3D-aware shadow and occlusion robust GAN for makeup transfer. In: Proceedings of the ACM Multimedia Conference on Multimedia Conference (ACM MM), pp. 3601–3609 (2021)
15. Xiang, J., Chen, J., Liu, W., Hou, X., Shen, L.: RamGAN: region attentive morphing GAN for region-level makeup transfer. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 719–735 (2022)
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), pp. 2672–2680 (2014)
17. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of International Conference on Computer Vision (ICCV), pp. 2242–2251 (2017)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695 (2022)
19. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention

control. In: Proceedings of International Conference on Learning Representations (ICLR) (2023)
20. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv Preprint arXiv:2204.06125 (2022)
21. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Phottorealistic text-to-image diffusion models with deep language understanding. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), pp. 36479–36494 (2022)
22. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22500–22510 (2023)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241 (2015)
24. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: Proceedings of International Conference on Learning Representations (ICLR) (2021)
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), pp. 6840–6851 (2020)
26. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: Proceedings of ACM SIGGRAPH, pp. 1–10 (2022)
27. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11451–11461 (2022)
28. Voynov, A., Aberman, K., Cohen-Or, D.: Sketch-guided text-to-image diffusion models. In: Proceedings of ACM SIGGRAPH, pp. 1–11 (2023)
29. Kwon, G., Ye, J.C.: Diffusion-based image translation using disentangled style and content representation. In: Proceedings of International Conference on Learning Representations (ICLR) (2023)
30. Liu, W., Liu, T., Han, T., Wan, L.: Multi-modal deep-fusion network for meningioma presurgical grading with integrative imaging and clinical data. Vis. Comput. 1–11 (2023)
31. Chang, H., Lu, J., Yu, F., Finkelstein, A.: PairedCycleGAN: asymmetric style transfer for applying and removing makeup. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 40–48 (2018)
32. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of International Conference on Machine Learning (ICML), pp. 2256–2265 (2022)
33. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: Proceedings of International Conference on Learning Representations (ICLR) (2021)
34. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. **23**, 47–14733 (2022)
35. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12873–12883 (2021)
36. Kim, G., Kwon, T., Ye, J.C.: DiffusionCLIP: text-guided diffusion models for robust image manipulation. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2416–2425 (2022)
37. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: SRDiff: single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
38. Liu, D., Wang, X., Peng, C., Wang, N., Hu, R., Gao, X.: Adv-diffusion: imperceptible adversarial face identity attack via latent diffusion model. arXiv preprint arXiv:2312.11285 (2023)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning (ICML), vol. 139, pp. 8748–8763 (2021)
40. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., Ermon, S.: SDEdit: guided image synthesis and editing with stochastic differential equations. In: Proceedings of International Conference on Learning Representations (ICLR) (2022)
41. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3836–3847 (2023)
42. Zhu, A., Lu, X., Bai, X., Uchida, S., Iwana, B.K., Xiong, S.: Few-shot text style transfer via deep feature similarity. IEEE Trans. Image Process. (TIP) **29**, 6932–6946 (2020)
43. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), pp. 8780–8794 (2021)
44. Nitzan, Y., Bermano, A., Li, Y., Cohen-Or, D.: Face identity disentanglement via latent space mapping. ACM Trans. Graph. **39**(6), 225–122514 (2020)
45. Afifi, M., Brubaker, M.A., Brown, M.S.: HistoGAN: Controlling colors of GAN-generated and real images via color histograms. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7941–7950 (2021)
46. Lan, J., Ye, F., Ye, Z., Xu, P., Ling, W.-K., Huang, G.: Unsupervised style-guided cross-domain adaptation for few-shot stylized face translation. Vis. Comput. 1–15 (2022)
47. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)
48. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 2555–2563 (2023)
49. Liu, D., Gao, X., Wang, N., Peng, C., Li, J.: Iterative local re-ranking with attribute guided synthesis for face sketch recognition. Pattern Recogn. **109**, 107579 (2021)
50. Liu, D., Gao, X., Peng, C., Wang, N., Li, J.: Universal heterogeneous face analysis via multi-domain feature disentanglement. IEEE Trans. Inf. Forensics Secur. **19**, 735–747 (2024)

Xiongbo Lu received the B.S. degree from the Hebei University of Technology, China, in 2014, and the M.S. degree from the Wuhan University of Technology, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Technology. His main research interests include image generation and style transfer.

Feng Liu received the B.S. degree in computer science and technology from the Anhui University, China, in 2014, and the M.S. degree in software engineering from the Nanjing University of Finance and Economics, China, 2018. He is currently pursuing the PhD degree with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology. His research interests include shape analysis, face analysis, and style transfer.

Yi Rong received the B.Sc. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2012, and the M.Sc. degree in software engineering and the PhD degree in computer science and technology from the Wuhan University of Technology, Wuhan, in 2014 and 2021, respectively. He is currently an associate professor with the School of Computer Science and Technology, Wuhan University of Technology, China. His research interests include machine learning, computer vision, pattern recognition.

Yaxiong Chen is an associate professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China. His research interests include pattern recognition, machine learning, hyperspectral image analysis, and medical imaging.

Shengwu Xiong received the B.Sc. degree in computational mathematics and the M.Sc. and PhD degrees in computer software and theory from Wuhan University, China, in 1987, 1997, and 2003, respectively. He is currently a professor with the School of Computer Science and Technology, Wuhan University of Technology, China. His research interests include intelligent computing, machine learning, and pattern recognition.