

High Fidelity Makeup via 2D and 3D Identity Preservation Net

JINLIANG LIU, University of Technology Sydney, Broadway, Australia

ZHEDONG ZHENG, University of Macau, Taipa, China

ZONGXIN YANG, Zhejiang University, Hangzhou, China

YI YANG, Zhejiang University, Hangzhou, China

In this article, we address the challenging makeup transfer task, aiming to transfer makeup from a reference image to a source image while preserving facial geometry and background consistency. Existing deep neural network-based methods have shown promising results in aligning facial parts and transferring makeup textures. However, they often neglect the facial geometry of the source image, leading to two adverse effects: (1) alterations in geometrically relevant facial features, causing face flattening and loss of personality, and (2) difficulties in maintaining background consistency, as networks cannot clearly determine the face-background boundary. To jointly tackle these issues, we propose the High Fidelity Makeup via two-dimensional (2D) and 3D Identity Preservation Network (IP23-Net), to the best of our knowledge, a novel framework that leverages facial geometry information to generate more realistic results. Our method comprises a 3D Shape Identity Encoder, which extracts identity and 3D shape features. We incorporate a 3D face reconstruction model to ensure the three-dimensional effect of face makeup, thereby preserving the characters' depth and natural appearance. To preserve background consistency, our Background Correction Decoder automatically predicts an adaptive mask for the source image, distinguishing the foreground and background. In addition to popular benchmarks, we introduce a new large-scale High Resolution Synthetic Makeup Dataset containing 335,230 diverse high-resolution face images to evaluate our method's generalization ability. Experiments demonstrate that IP23-Net achieves high-fidelity makeup transfer while effectively preserving background consistency. The code will be made publicly available.

CCS Concepts: • **Applied computing** → **Fine arts**; • **Computing methodologies** → **Computer vision representations**; **Image processing**;

Additional Key Words and Phrases: Makeup transfer, 3D stereoscopic loss, high resolution synthetic makeup dataset, multimedia application

ACM Reference Format:

Jinliang Liu, Zhedong Zheng, Zongxin Yang, and Yi Yang. 2024. High Fidelity Makeup via 2D and 3D Identity Preservation Net. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 8, Article 230 (June 2024), 24 pages. <https://doi.org/10.1145/3656475>

This work was supported by the National Natural Science Foundation of China (62293554) and the Fundamental Research Funds for the Central Universities (No. 226-2022-00051).

Authors' addresses: J. Liu, Faculty of Engineering and Information Technology, University of Technology Sydney, Broadway, New South Wales, Australia; e-mail: jinliang.liu@student.uts.edu.au; Z. Zheng, University of Macau, Taipa, China; e-mail: zhedongzheng@um.edu.mo; Z. Yang, Zhejiang University, Hangzhou, Zhejiang, China; e-mail: yangzongxin@zju.edu.cn; Y. Yang, Zhejiang University, Hangzhou, Zhejiang, China; e-mail: yangyics@zju.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 1551-6865/2024/06-ART230

<https://doi.org/10.1145/3656475>

1 INTRODUCTION

In recent years, more people have started sharing their selfie photos on social networks like Instagram and Facebook. Good-looking selfie photos can increase attractiveness and self-confidence to people. Although cosmetics can enhance one's appearance, they also prolong the time spent on makeup application [9]. Over the past few years, deep learning techniques have gained rapid momentum and are widely applied in various domains [6, 15, 16, 16, 29, 35, 37, 38, 42, 62, 63, 70]. This study focuses on an automatic makeup model that transfers makeup from a reference image to a source image while preserving the original identity and makeup style. However, most existing methods [2, 7, 10, 17, 21, 25, 27, 28, 36, 40, 43, 46, 60, 68] face challenges in preserving facial geometric information, distinguishing between facial foreground and background, and addressing the limitations of current makeup datasets. To tackle above challenges, we introduce the two-dimensional (2D) and 3D **Identity Preservation Net (IP23-Net)**, an end-to-end learning framework comprising a three-dimensional (3D) Shape Identity Encoder, a Makeup Style Encoder, and a Background Correction Decoder. As shown in Figure 1, we can observe the above problems clearly.

First, previous works often lose facial geometric information during makeup transfer, resulting in artificial appearances and the loss of the individual's distinctive features. This issue stems from existing methods not incorporating depth information and shadow representation, both crucial for the face's three-dimensional aspect. To address this challenge, we design a 3D Shape Identity Encoder consisting of a Shape Encoder, an Identity Encoder, and a Content Style Encoder. The Shape Encoder, based on a 3D face reconstruction model, introduces face depth information to enhance facial contours and makeup shadow perception, achieving a visually three-dimensional effect. In the Identity Encoder, a face recognition network extracts facial identity information, maintaining the distinctiveness of individual facial features even after makeup application and ensuring facial uniqueness preservation. Last, the Content Style Encoder extracts the face's style feature, supplementing detailed facial features. To augment the three-dimensional perception of the synthesized images, we propose a 3D stereoscopic loss that utilizes the 3D average face parameter derived from the **3D Morphable Model (3DMM)**. During this process, 3D face prior features are independently extracted from both the source and generated images using the 3DMM. Following this, the pixelwise L1 loss between the models is computed, effectively incorporating facial depth information into the generated image and achieving an enhanced stereoscopic effect. For makeup style extraction, our proposed Makeup Style Encoder obtains features from the reference image. In contrast to conventional encoders [27, 36], our approach shifts input focus from an entire image to specific facial regions based on the reference's corresponding mask. This enables IP23-Net to effectively perform makeup part selection and style blending.

Second, makeup transfer should differentiate between the facial foreground and background. However, most existing methods overlook this issue and modify background styles, leading to unrealistic artifacts. We gain inspiration from References [59, 69]. Our Background Correction Decoder generates the output image by fusing features from the 3D Shape Identity Encoder and Makeup Style Encoder and predicts an adaptive face mask using the source image's 3D structure feature. This mask explicitly delineates the boundary between the face's foreground and background, maintaining end-to-end characteristics and ensuring a realistic outcome.

Furthermore, the makeup transfer task lacks an effective validation dataset. The current state-of-the-art makeup MT dataset [36] comprises only 3,834 adult female facial images. As modern society evolves, the makeup user demographic has expanded from young women to various age groups and includes men. Consequently, makeup transfer techniques lack comprehensive evaluations for children, older adults, and male subjects. To address this issue, we introduce the **High-Resolution Synthetic Makeup (HRSM)** dataset, generated using StyleGAN2 [32] and comprising 335,230 facial images with diverse ages, poses, expressions, and backgrounds. This dataset enables a more

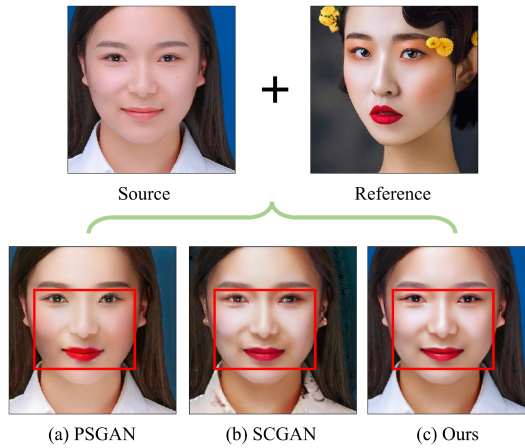


Fig. 1. IP23-Net is capable of generating highly stereoscopic images while accurately restoring the background. We compare our output image to the ones produced by state-of-the-art existing methods, PS-GAN [27] and SCGAN [10]. In addition to preserving the background, we also generate an output image featuring highly stereoscopic facial components, particularly in the mouth, nose, and eyes.

comprehensive evaluation of model generalization and accuracy. Notably, the HRSM dataset serves exclusively for evaluating the model's performance. Our IP23-Net method surpasses other state-of-the-art techniques in generating realistic images while effectively addressing existing methods' challenges. The main contributions of our work are as follows:

- We introduce, to the best of our knowledge, a novel makeup network utilizing a 3D face reconstruction model for high-fidelity makeup transfer, facial geometric information extraction, and face mask prediction to restore the background. A 3D stereoscopic loss based on 3DMM enhances the realism of the generated makeup.
- We present the HRSM dataset, comprising 335,230 diverse face images, which serves as the largest makeup dataset available for evaluating the generalization and accuracy of automatic makeup transfer models.
- Our extensive experimental results demonstrate that IP23-Net delivers competitive results in preserving facial identity and ensuring makeup transfer quality.

2 RELATED WORKS

2.1 Makeup Transfer

Makeup transfer aims to transfer a specific makeup style from a reference image to a source image. As one of the pioneering works, CycleGAN [72] investigates image style translation between unpaired images, which can be directly applied to the field of makeup. However, CycleGAN is designed for transferring global style between two pre-defined domains; thus, users are unable to customize the makeup style. Similarly to eyeglass removal [22], PairedCycleGAN [2] further refines the task of makeup. In this method, two translators are trained separately: one for facial makeup and another for facial makeup removal. Notably, the two translator structures are asymmetric. BeautyGAN [36] trains a semantic segmentation network for extracting masks from different facial regions, such as skin, eyes, and mouth, to calculate the makeup loss. Consequently, the color information of the makeup area can be restored in the source image. BeautyGlow [7] employs a generative flow model to obtain latent vectors from a source image and a reference image. The latent vectors are decomposed into facial features and makeup features, respectively.

Then, the makeup feature and the facial feature are combined to obtain the source image latent vectors with makeup. Ultimately, Chen et al. [7] apply the Glow model to invert the latent vectors to the RGB source image with makeup. Local Adversarial Disentangling Network [17] leverages multiple overlapping discriminators based on facial feature points for dramatic makeup transfer. PSGAN [27] proposes the **Attention Mechanism Module (AMM)** to explicitly address the spatial misalignment problem. Specifically, the AMM module aims to establish pixelwise correspondences between source images and reference images to align faces. SCGAN [10] draws inspiration from StyleGAN [31] to map makeup styles into an intermediate style space instead of using linearly projected vectors. The research by Lyu et al. [44] details a new technique, 3DAM-GAN, that effectively shields facial images from unauthorized detection by recognition systems. This innovation not only obscures personal identity but also proves to be highly effective in evading diverse facial recognition algorithms. Additionally, it also yielded satisfactory results in the task of makeup transfer. Although the aforementioned methods can achieve makeup transfer, they fail to restore the face depth information and background of the generated images. In contrast, our IP23-Net model can generate highly stereoscopic makeup images and accurately restore the non-makeup area.

2.2 Style Transfer

In recent years, style transfer has emerged as a prominent research area. The underlying concept of image style transfer involves extracting content and style features separately from source and target images and subsequently recombining these features to reconstruct the source image with the style of the target image. Certain methods [54, 57, 58, 71], aim to transfer a source image to various styles, such as plain and cartoon styles. Nonetheless, these methods are limited to transferring the source image to a detailed style.

2.3 3D-aware Image Synthesis

In prior research, numerous studies have focused on multi-view and 3D perceptual networks to enhance the fidelity in the image generation task [14, 19, 23, 48, 50, 53, 73], which is close with the temporal consistency in video generation [56] and multi-round editing [64]. HoloGAN [47] learns 3D features from rigid-body transformations to control the pose of generated objects. Schwarz et al. introduce GRAF [52], a method for high-resolution 3D-aware image synthesis that trains the model solely from unposed 2D images. PIGAN [1] achieves view-consistent, camera angle-controlled, high-quality image generation by employing neural representations with periodic activation functions and neural radiance field rendering. GIRAFFE [49] incorporates a compositional 3D scene representation into the generative model, improving the controllability of generated images, while Zhang et al. [66, 67] further consider the pose consistency.

3 APPROACH

In this section, we introduce our 2D and 3D IP23-Net, including Generator (G) and Discriminator (D). Given the source image x and the reference image y , our target is to learn a makeup transfer function $\hat{x} = G(x, y)$, where $\hat{x}$ preserves the identity of the source image x and the makeup style of the reference image y . Figure 2 shows an overview of our network and the detail of Generator. Our pipeline consists of three parts: 3D Shape Identity Encoder (E_{3D}), Makeup Style Encoder (E_M), and Background Correction Decoder.

3.1 3D Shape Identity Encoder

The E_{3D} is a vital component in our proposed makeup transfer framework, IP23-Net, designed to preserve the geometric structure and facial identity information of the source image. By

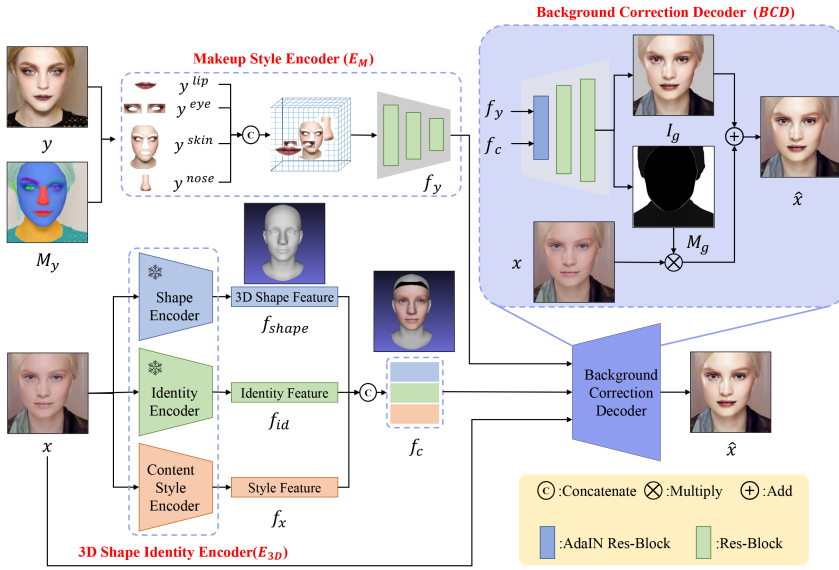


Fig. 2. Illustration of our IP23-Net workflow. Our pipeline comprises three components: the 3D Shape Identity Encoder E_{3D} , Makeup Style Encoder E_M , and Background Correction Decoder BCD. The 3D Shape Identity Encoder is constructed from a 3D face reconstruction model [13], a face recognition network [12], and a Content Style Encoder. These elements extract features f_{shape} , f_{id} , and f_x from the source image x , forming f_c . The Makeup Style Encoder retrieves style features f_y from the reference image y . Finally, the Background Correction Decoder fuses these feature types to generate the source face with the reference makeup and restores the background.

incorporating 3D information, we aim to overcome the limitations of existing makeup transfer methods that often fail to maintain the fidelity of facial geometry and identity. The E_{3D} module is comprised of three key parts: a Shape Encoder, an Identity Encoder, and a Content Style Encoder. To enhance the robustness of our model, we utilize pre-trained Shape Encoder and Identity Encoder with frozen parameters during network training. This approach allowed us to leverage prior knowledge and optimize the network for specific tasks, without extensive training from scratch.

Shape Encoder. The Shape Encoder is designed based on a state-of-the-art 3DMM [13] that extracts the facial geometry information from the source image. The facial geometry information consists of shape, texture, and depth information from the source image x . Depth information is of particular importance, as it provides rich geometric information such as contours, shapes, and local details. We define the facial geometry information extracted from the source image x as f_{shape} .

Identity Encoder. The process of 3DMM for facial modeling typically involves fitting the input face to the 3D average face model in the **Basel Face Model (BFM)** database [51] to predict the face model of the input image. However, since the face texture information of the predicted 3D face model is also sourced from BFM, the original details of the input image cannot be fully accurately. To address this issue and retain the identity information of the source image, we propose the use of a face recognition network [12] to extract the face identity feature f_{id} from the source image x . By using a face recognition network, we can better capture the individual differences and details in the input image, ensuring that the generated face model is consistent with the input image.

Content Style Encoder. For the makeup transfer, capturing not only the shape and identity features but also the style information of the face, such as skin color and lip color, is essential for

generating realistic and natural-looking images. To supplement this information and incorporate it into the generated images, we introduce the Content Style Encoder. It is designed to extract and encode the style information f_x from the source image x . The Content Style Encoder is comprised of multiple seven res-blocks and five downsampling convolutional layers, which enable the extraction and encoding of the style information in the source image.

To generate the final makeup transferred image, we concatenate f_{shape} , f_{id} , and f_x and feed them into the decoder. However, since f_{shape} and f_{id} are obtained using pre-trained models, their feature shapes cannot match that of f_s . To ensure consistency, we adjust the feature shapes of f_{shape} and f_{id} to match that of f_x using a 1×1 convolutional layer to adjust the number of channels, and a transposed convolutional layer to adjust the spatial resolution of the features. The final output of the 3D Shape Identity Encoder is denoted as fc , and the function is defined as

$$fc = \text{Concat}(\text{Conv}_{shape}(f_{shape}), \text{Conv}_{id}(f_{id}), f_x). \quad (1)$$

3.2 Makeup Style Encoder

The Makeup Style Encoder is designed to extract features from the reference image using an encoder-bottleneck architecture. While most existing methods obtain style codes by simply averaging facial features in a reference image, numerous makeup styles exist for various facial components, inevitably leading to entanglements between different facial makeups. In contrast to traditional approaches [27, 36], our Makeup Style Encoder adopts the SCGAN [10] strategy.

The first processing step of the Makeup Style Encoder decomposes each reference face into four parts (lips, skin, eyes, and nose) using the face parser [34] as follows:

$$y^i = y \odot S_{y,i}. \quad (2)$$

In this step, we process each component of the reference face as y^i , where $i = \{lip, skin, eyes, nose\}$. $S_{y,i}$ represents the corresponding mask of the reference image, and $\odot$ denotes the Hadamard product. After the initial processing, we concatenate these codes as input $Y = [y^{lip}, y^{skin}, y^{eyes}, y^{nose}]$ for E_M . The Makeup Style Encoder consists of multiple seven res-blocks and five downsampling convolutional layers, facilitating the extraction and encoding of style information from the reference image. Ultimately, we acquire the localized makeup style features, represented as f_y .

3.3 Background Correction Decoder

Our Background Correction Decoder reconstructs the source image with makeup, using the content feature f_c and the style feature f_y . Specifically, our decoder comprises eight res-blocks and five upsampling convolutional layers. The **Adaptive Instance Normalization (AdaIN)** layer [24] achieves style transfer by altering the data distribution of the feature map. We incorporate AdaIN layers [24] and two upsampling convolutional layers in the first five res-blocks to facilitate makeup transfer. The AdaIN layer is defined as follows:

$$\text{AdaIN}(f_c, f_y) = \sigma(f_y) \frac{f_c - \mu(f_c)}{\sigma(f_c)} + \mu(f_y), \quad (3)$$

where $\sigma(f_c)$ and $\mu(f_c)$ denote the mean and standard deviation of the source image features, respectively. $\sigma(f_y)$ and $\mu(f_y)$ represent the mean and standard deviation of the style image features, respectively. In general, AdaIN performs makeup transfer by transforming the mean and variance of specific channels in the feature map. We then employ another three res-blocks and upsampling convolutional layers to further fuse the feature maps and generate higher-resolution results. Note that these three res-blocks are all standard residual blocks without AdaIN layers. We incorporate the source image x as a residual to the output of the final residual blocks, which enhances the image restoration effect. Our approach is to ensure a balance between style transfer and content

preservation. While AdaIN layers excel in transferring makeup style, using them excessively might compromise content information. To avoid this, we limit the number of AdaIN layers. Res-blocks without AdaIN layers play a significant role in refining and blending features after the primary makeup transfer process.

Additionally, the background information of the generated image is susceptible to damage by the reference image style during makeup transfer. To better preserve the background, we use the decoder to learn an adaptive mask and accommodate the change in face shape. Specifically, we predict a three-channel I_g and one-channel M_g after the upsampling. The M_g helps us identify the facial foreground and background areas. We apply the generated face I_g and blend the background of the clean source image x to create the makeup image using M_g , formulated as

$$\hat{x} = I_g \odot M_g + (1 - M_g) \odot x, \quad (4)$$

where $\hat{x}$ denotes the generated makeup image. Despite changes in face shape, this decoder remains valid because the predicted mask M_g is deformable to follow the input source image.

4 OPTIMIZATION OBJECTIVES

Our IP23-Net aims to learn the corresponding relationships between the source image x and the reference image y .

4.1 3D Stereoscopic Loss

Traditional approaches apply 2D landmark keypoints as geometric supervision to constrain face shape [27]. To generate highly stereoscopic images, we introduce a 3D face reconstruction model for obtaining facial depth information. Specifically, we use a pre-trained, state-of-the-art 3DMM [13] to accurately extract facial geometric information. First, we employ the regressing 3DMM coefficients model [13] to obtain the 3D face model and calculate the L1 loss as follows:

$$L_{attribute}^{3D} = ||f_{shape}^x - \hat{f}_{shape}^x||_1 + ||f_{shape}^y - \hat{f}_{shape}^y||_1, \quad (5)$$

where $f_{shape}^x, \hat{f}_{shape}^x$ denote the 3D face attributes of the source image and generated makeup image and $f_{shape}^y, \hat{f}_{shape}^y$ represent the 3D face attributes of the reference image and generated non-makeup image. Note that $||\cdot||_1$ is the L1-norm loss function. Second, we project the 3D facial vertices from the 3D model onto the 2D image to obtain landmarks. We then calculate the L1 loss between these landmarks to enhance the effect of the 3D stereoscopic loss as follows:

$$L_{landmark}^{3D} = ||k^x - \hat{k}^x||_1 + ||k^y - \hat{k}^y||_1, \quad (6)$$

where $k^x, \hat{k}^x$ denote the 2D landmarks of the source image and generated makeup image, while $k^y, \hat{k}^y$ represent the 2D landmarks of the reference image and generated non-makeup image. Finally, our 3D stereoscopic loss is formulated as follows:

$$L_{3D} = \lambda_{attribute}^{3D} L_{attribute}^{3D} + \lambda_{landmark}^{3D} L_{landmark}^{3D}, \quad (7)$$

where $\lambda_{attribute}^{3D} = 2$ and $\lambda_{landmark}^{3D} = 0.2$.

4.2 Adversarial Loss

The adversarial loss is applied to generate more realistic results. We employ discriminator D_x , which attempts to distinguish between fake and real makeup images. Additionally, we propose discriminator D_y to differentiate between fake and real non-makeup images. Discriminators D_x and D_y classify 70×70 image patches using the same architecture as PatchGANs [26] and do not

share parameters. The adversarial losses L_{D_x} and L_{D_y} are defined as follows:

$$\begin{aligned} L_{D_x} &= \mathbb{E}[\log D_x(y) + \log(1 - D_x(G(x, y)))], \\ L_{D_y} &= \mathbb{E}[\log D_y(x) + \log(1 - D_y(G(y, x)))]. \end{aligned} \quad (8)$$

Finally, the adversarial loss for the Generator (G) integrates the two terms L_{D_x} and L_{D_y} as follows:

$$L_{adv} = -L_{D_x} - L_{D_y}. \quad (9)$$

Note that the tasks of G and D_x, D_y are reversed.

4.3 Perceptual Loss

Our perceptual loss includes the global perceptual loss and the local perceptual loss. The global perceptual loss [30] aims to preserve personal identity information. Specifically, we utilize the VGG-16 [55] model pre-trained on ImageNet [11] to compare the activation of the source and generated images in the hidden space, L_{global} is defined as follows:

$$L_{global} = \|V(G(x, y)) - V(x)\|_2 + \|V(G(y, x)) - V(y)\|_2, \quad (10)$$

where $V(\cdot)$ is defined as the output of the VGG-16 model and $\|\cdot\|_2$ is the L2-norm loss function. Besides the global perceptual loss, we introduce the local perceptual loss [10] to preserve the non-makeup parts. Specifically, the local perceptual loss is the same as the global perceptual loss but changes the input from a global image to the non-makeup facial component parts, L_{local} is defined as follows:

$$\begin{aligned} L_{local} &= \sum_{n=1}^N [\|V(G(x, y) \odot S_{x,n}) - V(x \odot S_{x,n})\|_2] \\ &\quad + \sum_{n=1}^N [\|V(G(y, x) \odot S_{y,n}) - V(y \odot S_{y,n})\|_2], \end{aligned} \quad (11)$$

where $S_{x,n}$ and $S_{y,n}$ are the facial component mask of the source image and the reference image, respectively. The unchanged non-makeup components $N = \{hair, eyeballs, teeth\}$, n denotes the index of N . The final perceptual loss is defined as follows:

$$L_{per} = L_{global} + L_{local}. \quad (12)$$

4.4 Cycle Consistency Loss

We utilize the cycle consistency loss [72] to train our network in an unsupervised way. If we translate a non-makeup image to the makeup domain and back, then the cycle consistency loss [72] helps us to reconstruct this non-makeup image. The cycle consistency loss is pixelwise supervision to support our network produce high-fidelity results, L_{cyc} is defined as follows:

$$L_{cyc} = \|G(G(x, y), x) - x\|_1 + \|G(G(y, x), y) - y\|_1, \quad (13)$$

where $\|\cdot\|_1$ denotes the L1-norm loss function.

4.5 Makeup Loss

The makeup loss [36] applies histogram matching to ensure that the generated makeup image and the reference image have the same makeup color. We apply histogram matching on the same facial regions (lip, eyes, skin, and nose) from x and y , which are extracted by the face masks [34] S_x and S_y , respectively, and recombine them as $HM(x, y)$. The $HM(x, y)$ retains the content information of the source image x and the makeup color distribution of the reference image y for a coarse guidance. Finally, we calculate the L2 loss between $HM(x, y)$ and the generated makeup image $\hat{x}$

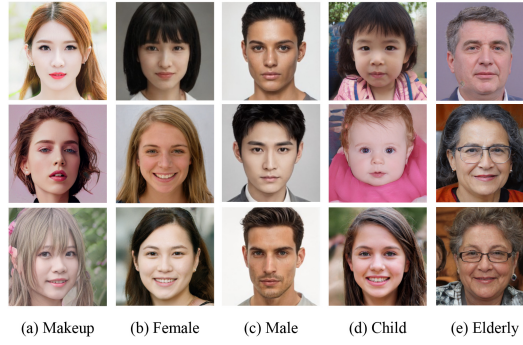


Fig. 3. Samples from the proposed HRSM dataset. The images displayed from (a) to (e), depict makeup faces, females, males, children, and elderly individuals.

for each channel of the RGB image. In addition, we also calculate the makeup loss between the generated non-makeup image $\hat{y}$ and $HM(y, x)$. The L_{HM} is defined as follows:

$$L_{HM} = \|\hat{x} - HM(x, y)\|_2 + \|\hat{y} - HM(y, x)\|_2, \quad (14)$$

where $\|\cdot\|_2$ denotes the L2-norm loss function.

Total Loss. The total loss L_{total} of our network is defined as follows:

$$L_{total} = L_{3D} + L_{adv} + L_{per} + L_{cyc} + L_{HM}. \quad (15)$$

5 HIGH-RESOLUTION SYNTHETIC MAKEUP DATASET

We observe that the sample number of datasets YMU [4], VMU [8], MIW [3], and MIFS [5] is less than 1,000 (see Table 1). The **Makeup Transfer (MT)** dataset [36] further enlarges the dataset size to 3,834, which contains 1,115 non-makeup images and 2,719 makeup images. Most existing approaches train and test models on the MT datasets. However, it only includes adult female samples with a relatively low resolution (361×361). As a result, the MT dataset lacks diverse samples (e.g., males and kids) to validate the generalizability of the model. To overcome the data limitation, we propose a High Resolution Synthetic Makeup dataset. First, we deploy StyleGAN2 [32] model to generate 500,000 face images with various races, poses, expressions, and backgrounds, and the resolution is 1024×1024 . Second, we align the generated images with 68 landmarks. Last, we obtain the gender and age labels through **Open Source Computer Vision (OpenCV)** libraries. We utilize the MT dataset [36] to train a binary classifier to distinguish makeup images and non-makeup images. In this process, we use different pre-trained models such as VGG-19 [55], ResNet-50 [18], and Swin Transformer [41]. Then, we ensemble the classification results of the three models together to increase the accuracy of the labels. Most of the images with the correct classification results. To further improve the label accuracy in the HRSM dataset, we set a confidence interval scope to remove images with low confidence. As a result, we obtained 225,831 makeup images and 109,399 non-makeup images for a total of 335,230 images. The examples of the HRSM dataset are illustrated in Figure 3. Figure 4 demonstrates the proportion and number of each label in our dataset.

To the best of our knowledge, the HRSM dataset is the first makeup transfer dataset with 1024×1024 resolution and contains 335,230 samples. The advantages of our collection method are scalability and security. Scalability means our dataset can be expanded conveniently. Specifically, we can harness StyleGAN2 [32] model to generate more face images and then use our classifier to distinguish the makeup images and non-makeup images. For the advantage of security, the face images we generated by StyleGAN2 [32] are not real-world images. Therefore, people using the

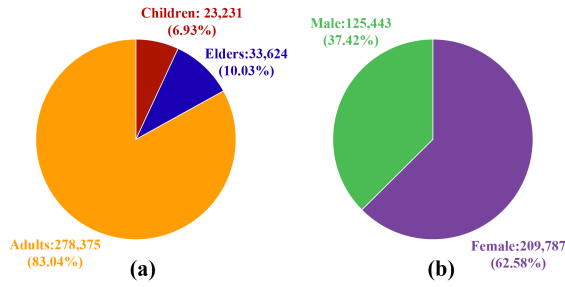


Fig. 4. Attribute distribution of the proposed HRSM dataset. (a) Percentage of children, adults, and elders. (b) Percentage of male and female.

Table 1. Comparison with Other Public Makeup Datasets

Dataet	ID number	Total number of images	Age label	Gender label
YMU [4]	151	604		
VMU [8]	51	204		
MIW [3]	125	154		
MIFS [5]	214	642		
MT [36]	3,000+	3,834		
Ours	330,000+	335,230	✓	✓

HRSM dataset are unable to violate the privacy of other people. The detailed comparison between makeup datasets is in Table 1.

6 EXPERIMENTS

6.1 Implementation Details

For a fair qualitative comparison, we train IP23-Net on the MT dataset [36]. The input images are resized to 256×256 for the low-resolution setting. We use the *relu_4_1* feature layer of VGG-16 [55] network to calculate the perceptual loss. The weights of different loss functions are set as $\lambda_{3D} = 1$, $\lambda_{adv} = 10$, $\lambda_{global} = 0.005$, $\lambda_{local} = 0.05$, $\lambda_{cyc} = 10$, and $\lambda_{HM} = 1$. Our network is optimized by Adam [33] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is 0.0002 and the batch size is set to 1.

6.2 Qualitative Comparisons

We compare our IP23-Net with other image-to-image makeup translation methods, including DIA [39], CycleGAN [72], PairedCycleGAN [2], BeautyGlow [7], LADN [17], BeautyGAN [36], PSGAN [27], and SCGAN [10] as well as two recent makeup transfer methods CPM [45] and EleGAN [61]. Figure 5 demonstrates the qualitative comparison of IP23-Net with other methods on the generated makeup images. The results produced by DIA [39] show the abnormal color of the hair, and the color of the lipstick is not transferred well. In addition, the background information is affected by the reference image. Although CycleGAN [72] can be leveraged to transfer makeup, the results are incomplete. BeautyGAN [36] performs well on the makeup transfer. Nevertheless, we can find unnatural color changes in the hair part between the generated and source images. Besides, BeautyGlow [7] transfers the makeup to the source images with an uneven color distribution in the lipstick part. We use the released model from LADN [17] to obtain the generated makeup

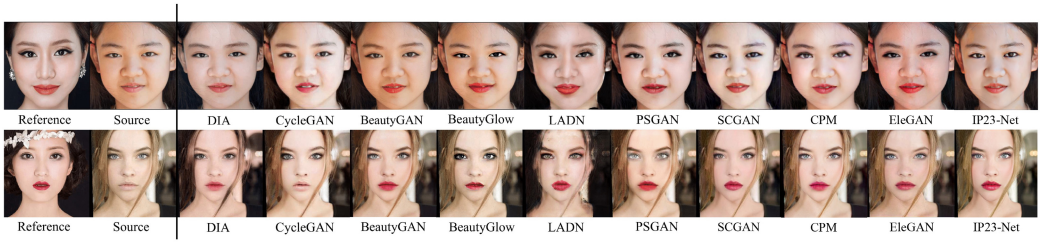


Fig. 5. Qualitative comparison with existing models. Both source and reference images are selected from the MT dataset for a fair comparison. IP23-Net effectively transfers the makeup style from the reference image to the source image. In addition, our method preserves the background as well as the original identity.

images. The background of the generated images is blurred, and some uneven color blocks appear on the face. PSGAN [27] produces fewer artifacts in the generated images. However, there is a problem of inaccurate color restoration. Specifically, the color of the lip between the generated and reference images is inconsistent. SCGAN [10] can address the spatial misalignment problem. For the CPM [45], we note that although it is able to achieve more satisfactory makeup transfer in specific scenarios, it may be slightly limited when dealing with diverse makeup style transitions. As for EleGAN [61], it performs well in localised makeup editing, but there is still room for further improvement in the naturalness and detailing of full-face makeup. The above limitations can be summarised in three points. First, the shadow information is lost in the generated image, resulting in an unrealistic nose. Second, the texture information is not preserved well. For example, the size of the lip in the generated image is smaller than the source image. Third, the background and the hair color of the generated images are changed. Compared to these methods above, our IP23-Net can generate a high-fidelity face and preserve the background accurately.

We evaluate the realism of images generated by IP23-Net using the **Fréchet Inception Distance (FID)** [20], comparing it against BeautyGAN [36], PSGAN [27], and SCGAN [10] under uniform conditions. We select makeup and non-makeup images from the MT dataset [36], transforming non-makeup images into a consistent makeup style. Different reference images are used to calculate the average FID. IP23-Net achieves the lowest FID score of 45.59, indicating superior image quality. To validate the identity preservation when performing makeup transfer, we use ArcFace [12] to calculate the similarity between the faces before and after makeup transfer. Comparison tests with other methods indicate that IP23-Net obtained the highest score of 0.9743, demonstrating better feature maintenance. Additionally, we incorporate the **Natural Image Quality Evaluator (NIQE)** to gauge the visual quality of the images. NIQE is a non-referenced measure that assesses the naturalness of an image based on the statistical properties of a natural, undistorted image. As Table 2 shows, our IP23-Net scores 5.6377 in NIQE, outperforming others in visual naturalness and reducing perceptual distortion, thus confirming the effectiveness of our approach in maintaining image quality and fidelity.

To assess the preservation of expression and pose, we employed a pre-trained MTCNN model to detect facial landmarks in source images. We then calculated the cosine similarity (CosSim $\in [0, 1]$) of these landmarks to evaluate our method's efficacy in maintaining facial expressions and poses. As Table 3 shows, IP23-Net and other baseline methods demonstrate effective preservation of expressions and poses.

6.3 Computational Complexity Analysis

For the tests regarding the computational complexity of our IP23-Net, we also compared it to previous approaches, and for a fair comparison, we used a single RTX6000 GPU uniformly in our tests, and all models are evaluated using the test section of the MT dataset to ensure fairness.

Table 2. Comparison between IP23-Net and Other Competitive Methods in Terms of FID, NIQE (Lower Is Better), and Arcface (Higher Is Better)

Method	IP23-Net	SCGAN[10]	PSGAN[27]	BeautyGAN[36]
FID ↓	45.59	47.78	51.78	57.43
Arcface ↑	0.9743	0.9611	0.9721	0.9691
NIQE ↓	5.6377	6.1762	5.7311	5.8729

The best results are highlighted in bold.

Table 3. Our Method Protects Expressions/Poses by Computing the Cosine Similarity between Landmarks

Method	IP23-Net	SCGAN[10]	PSGAN[27]	BeautyGAN[36]
CosSim ↑	0.9994	0.9971	0.9992	0.9992

The best result is highlighted in bold.

Table 4. Comparison of Time Efficiency of our Method with Other Competitors

Methods	BeautyGAN	PSGAN	SCGAN	Ours (IP23-Net)
Time (sec./image)	0.0156	1.1586	0.1272	0.0241

Table 4 illustrates that IP23-Net significantly outstrips competitors with an average inference speed of only 0.0241 seconds per image. This compares favorably to the times recorded by PSGAN and SCGAN. Such efficiency is maintained even with the inclusion of complex pre-trained models like the 3D Morphable Model and Arcface in our encoder framework, which are highly optimized for feature extraction and impose minimal computational burden. IP23-Net thus offers an optimal blend of rapid inference and sturdy performance, essential for real-time processing scenarios.

6.4 Partial Makeup Transfer

IP23-Net supports partial makeup transfer. People are able to choose the area of makeup by themselves from different reference images. Given a source image x and two reference images y_1 and y_2 . We can use the lip part from y_2 and other parts from y_1 to obtain the new mixed makeup area $Y = [y_1^{skin}, y_1^{eyes}, y_1^{nose}, y_2^{lip}]$, where $[\cdot, \cdot]$ denotes concatenation operation. The input to Facial Component Style Encoder is the facial components, and we can select specific parts from any reference image. The partial makeup transfer results are presented in Figure 6. We can find that the style of the lip in the generated image is from Reference2, and the others (eyes, nose, and skin) are from Reference1.

6.5 Shade-controllable Makeup Transfer

Our model can control the degree of makeup easily. We use a style-code f_{all} to represent the input feature of Background Correction Decoder, consisting of content feature f_c and makeup style feature f_y . We set a coefficient $\alpha \in [0, 1]$ to control the weight of f_c and f_y , which is defined as $f_{combined} = \alpha f_y + (1 - \alpha) f_c$. Figure 7(a) shows the makeup degree from light to heavy. We can see that the makeup degree of the source image gets higher when the coefficient α increases. In addition, IP23-Net supports the makeup style fuse from multiple reference images. We can also use the coefficient α to change the makeup style of the source image from Reference1 to Reference2. The style features f_{y_1} and f_{y_2} are from Reference1 and Reference2, respectively. The style-code f_{all} is calculated by $f_{all} = (1 - \alpha) f_{y_1} + \alpha f_{y_2}$. Figure 7 (b) shows the makeup style changes from

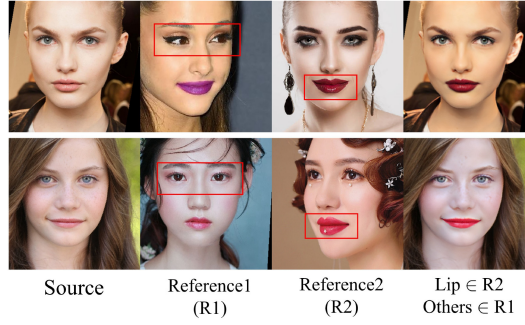
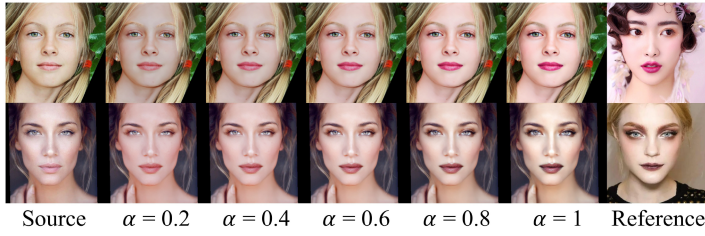


Fig. 6. Results of partial makeup transfer. The lipstick style for the generated image is from Reference2, and the other makeup styles are from Reference1. The source and reference images are from the MT dataset.

(a) Single Shade-controllable Makeup Transfer



(b) Multiple Shade-controllable Makeup Transfer

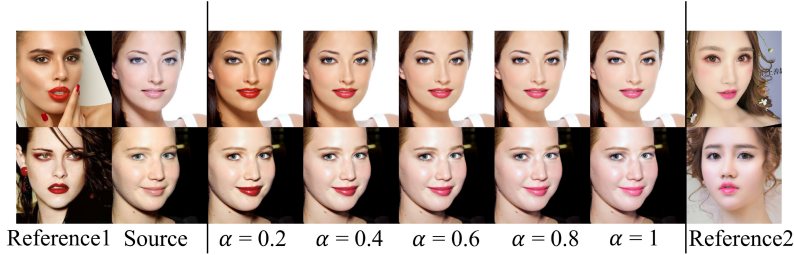


Fig. 7. (a) Shade-controllable makeup transfer results. The results are sorted from left to right according to the degree of makeup from light to heavy. (b) The makeup style of the generated images gradually changes from Reference1 to Reference2 from left to right. Note that the displayed source and reference images are from the MT dataset.

Reference1 to Reference2 when the coefficient α increases. In other words, the makeup style is closer to the reference image that contributes more style feature to f_{all} .

6.6 Generalization to Unseen Images

In this study, we introduce, to the best of our knowledge, the novel High Fidelity Makeup Transfer Network via 2D and 3D Identity Preservation, termed as IP23-Net. For test the generalisability of our IP23-Net. We train the model on the publicly available MT dataset [36] and test on the unseen images from HRSM Dataset. Figure 8 demonstrates the efficacy of our method. Moreover, to comprehensively assess the robustness of our model, we selected unseen male images from the HRSM dataset for additional testing. Figure 9 illustrates the results of our makeup application on male subjects. To further validate the performance and generalizability of our proposed IP23-Net



Fig. 8. Results of makeup transfer. The source and reference images are from the HRSM (ours) dataset.

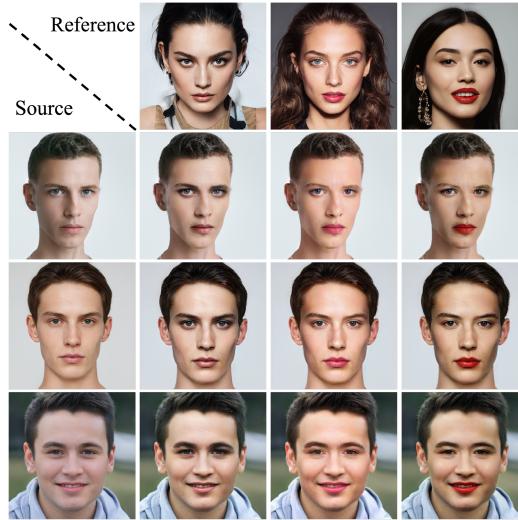


Fig. 9. Male makeup results. The source and reference images are from the HRSM (ours) dataset.

method, we have also conducted visualizations of makeup transfer effects on the Flickr-Faces-HQ dataset [31] and EDFace-Celeb-1M dataset [65] (shown in Figure 10). In addition, to more fully demonstrate the performance of our method in the transfer of different makeup styles, we purposely add Figure 11 to visualise a randomly selected case. As you can see, the left part shows the results of a heavier makeup transfer, while the right part shows the results of a more traditional Asian makeup style transfer. By comparing the results of these two styles of makeup, we can see the flexibility and efficiency of IP23-Net in handling different makeup styles.

From all the above evidence, it is evident that our model not only delivers high-quality results across various makeup styles but also exhibits an impressive balance between transformation and preservation. Notably, the backgrounds of the characters remain intact after the makeup transfer, maintaining the overall integrity of the image.

6.7 Comparison of Facial Details

Figure 12 left part presents a detailed comparison of our IP23-Net with state-of-the-art methods PSGAN [27] and SCGAN [10]. The makeup images generated by our network exhibit notable

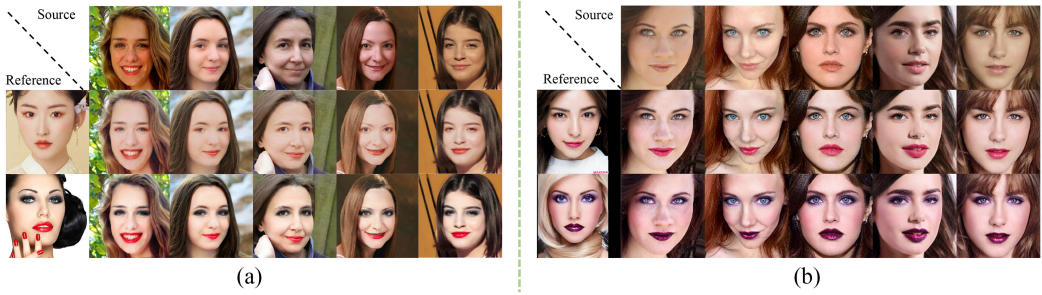


Fig. 10. Results of makeup transfer by IP23-Net. (a) The source and reference images are from the Flickr-Faces-HQ (FFHQ) dataset [31]. (b) The source and reference images are from the EDFace-Celeb-1M dataset [65].

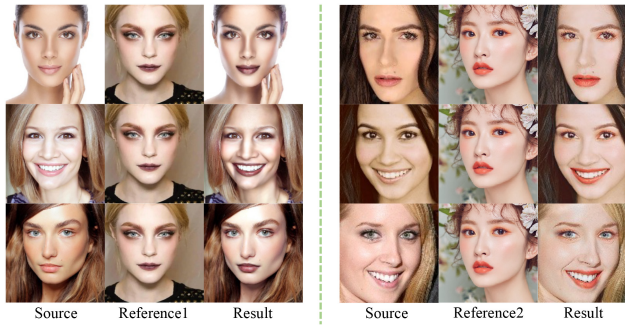


Fig. 11. Results of makeup transfer by IP23-Net. The different source images are selected randomly, two makeup styles with widely differing styles are selected as reference images.

superiority in retaining more character details and shadows, especially within the highlighted regions. This is a significant improvement over the prior methods, enhancing the three-dimensionality of the characters and making the makeup appear more realistic. A key feature of our approach is its ability to enhance these aspects without compromising the unique personality of the characters, thereby achieving a balance between transformation and authenticity that previous methods have struggled to attain. This balance results in a more holistic and appealing representation of makeup transfer, thereby pushing the boundaries of what is achievable in this field. Furthermore, our approach not only excels in executing high-quality makeup transfer but also demonstrates a strong commitment to maintaining the original image's integrity. This effectiveness is achieved by restoring the background of the characters and non-makeup areas, as illustrated in an ID photo presented in Figure 12 right part. This figure offers a detailed comparison with state-of-the-art methods PSGAN [27] and SCGAN [10]. In contrast to these previous methods, IP23-Net demonstrates high accuracy in reconstructing the character background, further preserving the authenticity of the original image. The content within the black box accentuates the superior capacity of IP23-Net for character detail preservation, enhancement of image three-dimensionality, and maintenance of original image integrity. The effectiveness of our approach in background restoration, combined with its proficiency in makeup transfer, marks a significant advancement in the field. It establishes a balance between aesthetic enhancement and preservation of the original image, ensuring a realistic and holistic representation in makeup transfer tasks.

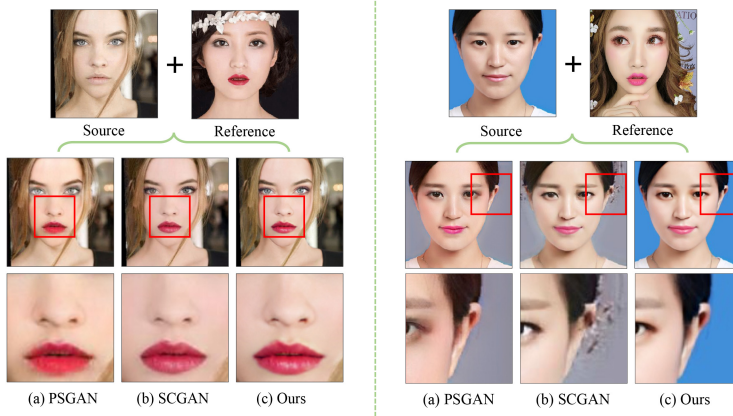


Fig. 12. Local detailed comparison. Both source and reference images are selected from the MT dataset [36]. From the left side of the figure, it is evident that IP23-Net is able to generate highly stereoscopic makeup images with accurate background restoration. By comparing the content in the black box, we found that our generated lipstick is plump and glossy. In addition, according to the right part of the figure, we can find out that our method is also satisfactory for background reconstruction.

Table 5. Evaluation Results of Different Makeup Transfer Methods in User Studies, Represented by the Ratio Selected as Best (RSB) in Percentage

Method	IP23-Net	SCGAN[10]	PSGAN[27]	BeautyGAN[36]
RSB $\uparrow$	53.68	16.55	24.72	5.05

The best result is highlighted in bold.

6.8 User Study

We conducted plausible user studies to quantitatively evaluate the robustness and visual quality of IP23-Net with three makeup transfer methods, BeautyGAN, PSGAN, and SCGAN. A total of 20 participants took part in these user studies. In the studies, we randomly selected 15 pairs of makeup and non-makeup images. The aim was to investigate the completeness of the makeup migration and whether the contours of the face were well preserved, in addition to whether the non-makeup areas of the characters were affected. In all user studies, participants were asked to select the result with the best visual quality and the most accurate transfer. Table 5 illustrates the results of the user studies, where our IP23-Net outperforms all state-of-the-art methods. We have reason to believe that this is due to the fact that we have introduced 3D face information reasonably well into the training process of the model, thus preserving the depth information of the face well.

6.9 Ablation Study

6.9.1 Impact of 3D Stereoscopic Loss on Image Quality. To verify the effect of different losses, we perform ablation studies on the MT dataset. Specifically, we first add losses one by one to train different IP23-Nets under the same experimental setting. We then calculate the FID [20] between the reference image and the generated makeup image. Table 6 demonstrates the impact of each loss function in terms of the performance. The baseline model, which employed only adversarial loss, has achieved an FID of 80.49. By incorporating cycle consistency loss, the FID significantly decreased to 60.99, showing the importance of preserving the subject's identity during makeup transfer. The addition of perceptual loss led to an FID reduction to 59.06, implying that high-level semantic information from pre-trained models contributes to the preservation of facial details and

Table 6. Ablation Study for Different Losses

Loss	Performance				
Adversarial Loss	✓	✓	✓	✓	✓
Cycle Consistency Loss		✓	✓	✓	✓
Perceptual Loss			✓	✓	✓
Makeup Loss				✓	✓
3D Stereoscopic Loss					✓
FID ↓	80.49	60.99	59.06	48.06	45.59

The best result is highlighted in bold.

makeup style. With the introduction of makeup loss, performance improved further, and the FID dropped to 48.06, emphasizing its effectiveness in preserving and transferring makeup attributes. Last, the inclusion of 3D stereoscopic loss resulted in the lowest FID of 45.59, signifying the value of 3D facial structure information for enhanced makeup transfer and more natural results. The ablation studies clearly demonstrate that the combination of adversarial loss, cycle consistency loss, perceptual loss, makeup loss, and 3D stereoscopic loss is crucial for achieving optimal performance in our makeup transfer network.

6.9.2 Fixed-parameter Encoder Performance Evaluation. In our IP23-Net architecture, the design of the 3D Shape Identity Encoder incorporates two distinct encoder components: the Shape Encoder and the Identity Encoder. The Shape Encoder uses a 3DMM [13], which exists in a parametric form and is primarily utilized for reconstructing 3D facial structures from 2D images. This model reconstructs the 3D facial structure from 2D images by comparing the input facial data with the average facial model in the BFM database. These parameters are pre-calculated based on extensive detailed 3D facial scan data and are not obtained through training. Therefore, in the Shape Encoder, we use a fixed 3DMM that accurately captures facial geometry, which is crucial for high-fidelity makeup transfer. However, the Identity Encoder employs the Arcface model [12], an advanced facial recognition network used to extract facial identity features. We added experiments to compare the differences between using the fixed-parameter Arcface model and an Arcface model with trainable parameters in the Identity Encoder. We compare two key performance metrics on the test set of the MT dataset: Arcface Similarity (higher is better) and FID (lower is better).

In Table 7, the fixed-parameter(*) Identity Encoder demonstrates superior performance over the trainable parameter configuration for identity consistency and image authenticity in makeup transfer tasks. This is evidenced by the reduced FID from 47.23 to 45.59, indicating enhanced efficiency, and the increase in similarity from 0.9743 to 0.9631, highlighting its effectiveness. The fixed-parameter approach, pre-trained on a diverse dataset, effectively captures facial features, ensuring high-fidelity in makeup transfer. Its constant parameters during training contribute to stable outputs across various makeup styles. This stability is crucial for accurately replicating reference makeup styles while preserving the source identity.

6.9.3 Effects of Shape Encoder. We conduct additional experiments, selecting the seminal work in makeup transfer, BeautyGAN, as the foundation for our study. On this basis, we integrate our innovative feature, the Shape Encoder, and employ 3D stereoscopic loss as the supervisory mechanism. In line with the method we describe in our article, we incorporate the features generated by the Shape Encoder (f_{shape}) into the network through concatenation. We then apply a 1×1 convolutional layer to adjust the number of channels, ensuring effective integration and functioning of the features. To illustrate the enhancement's impact more clearly, we provide a visual comparison

Table 7. Performance Comparison of IP23-Net with Fixed (*) and Trainable Identity Encoder Configurations in Terms of FID (Lower Is Better) and Arcface Similarity (Higher Is Better)

Configuration	FID ↓	Arcface ↑
IP23-Net (Identity Encoder)	47.23	0.9631
IP23-Net (* Identity Encoder)	45.59	0.9743

The best results are highlighted in bold.

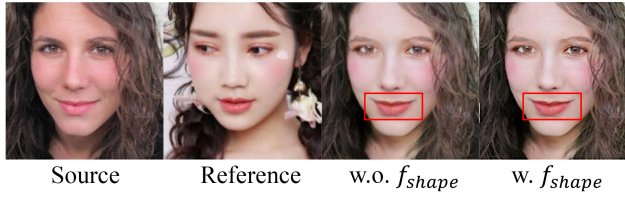


Fig. 13. Comparative Visualization of Makeup Transfer with and without the Shape Encoder.

that contrasts with the original BeautyGAN method in Figure 13. By observation, we can see that although the overall difference between the two generated makeup images is not very large, if we look closely at the lip part in the black box, then we can still feel that the resulting image with the addition of f_{shape} has a more defined facial contour.

Finally, we believe that despite the integration of the Shape Encoder in BeautyGAN, there are some limitations in handling the interaction of f_{shape} features as the framework does not include our specially designed Background Correction Decoder. Due to insufficient feature fusion, the potential of the Shape Encoder could not be fully utilised, resulting in a less than obvious enhancement of the visualisation. Based on these observations, we conclude that Shape Encoder can be useful in different approaches, but its value can only be maximised in the IP23-Net framework.

6.9.4 Evaluating the Impact of HRSM Dataset. Integrating the HRSM dataset with the MT dataset is expected to improve our model's robustness by diversifying its training data. The HRSM dataset, with its high-resolution and varied synthetic facial images, complements the MT dataset's style and complexity. This integration aims to broaden the model's understanding of diverse facial features and makeup styles, potentially improving its adaptability to complex real-world scenarios and its accuracy in makeup style transfer.

To test this hypothesis, we conducted a visual experiment, comparing models trained on the MT dataset alone with those trained on both MT and HRSM datasets, focusing on real-world image processing. The comparative results are illustrated in Figure 14. Crucially, all test images in our experiment originate from real-world scenarios and are not part of the model's training set, ensuring unbiased performance evaluation. The comparative analysis revealed that models trained with the HRSM dataset more accurately replicated the makeup styles from reference images while preserving the identity of the source images. This finding supports the notion that integrating the HRSM dataset into training enhances the model's performance in real-world applications.

6.9.5 Impact of Face Shape Information on Image Quality. In our study, we extracted facial depth information (f_{shape}) from the source images using a shape encoder. This is intended to enhance the three-dimensionality of the facial contours while preserving the original identity features of the subjects. An ablation study to validate the effectiveness of depth information in enhancing

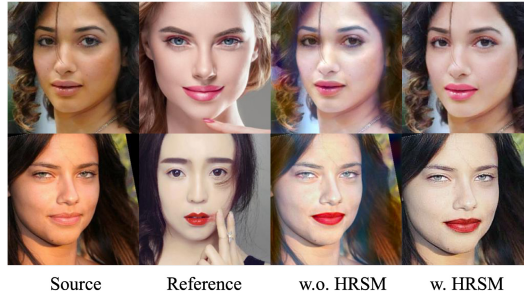


Fig. 14. Comparative results of makeup transfer models trained without (w.o. HRSM) and with (w. HRSM) the HRSM dataset, using real-world source and reference images.

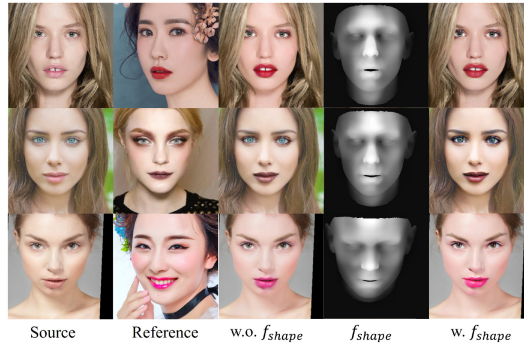


Fig. 15. This figure shows makeup transfer with (w.) and without (w.o.) facial shape feature f_{shape} . The source and reference images are from the MT dataset. We can observe that faces with f_{shape} exhibit a stronger sense of three-dimensionality in the nose and lip areas.

the three-dimensionality of makeup and preserving identity features is conducted. The results, as shown in Figure 15, demonstrate that models utilizing f_{shape} perform better in maintaining facial contours and identity features. This proves the significant role of integrating depth information in improving the overall performance and output quality of the makeup transfer process. This discovery provides intuitive insights into the critical components of our method and their impacts.

Discussion. The importance of depth information cannot be ignored in the process of makeup transfer. It allows makeup to be applied in a way that respects the natural profile and height of the face. This means that makeup is not just adding layers of color to the face, but takes into account the three-dimensional structure of the face, such as the height of the bridge of the nose, the hollows of the eye sockets, and the curves of the cheeks. Such a makeup transfer not only looks more natural, but also better adapts to the unique facial features of different individuals.

A key reason we chose to use 3DMM is its ability to efficiently apply this 3D depth information in a 2D image. Despite the fact that the final output is on a 2D plane, the depth information obtained through 3DMM can enhance the three-dimensionality and dynamics of facial features. This approach overcomes the limitations in traditional 2D image processing, making makeup transfer not just a simple migration of colors, but a more comprehensive artistic creation that takes into account the 3D structure of the face. Therefore, 3DMM plays a crucial role in our makeup transfer process. It not only provides precise information about the depth of the face, but also ensures a

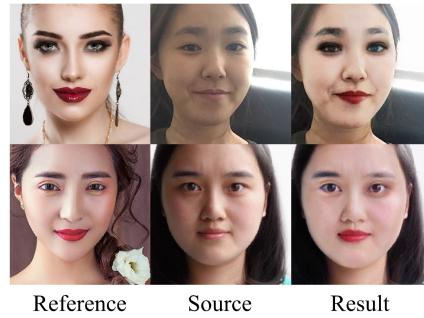


Fig. 16. Eyebrow Rendering Limitations: Although our method successfully transfers basic makeup, it struggles to accurately depict eyebrow special effects in scenarios with extensive facial shadows.

natural and personalized makeup effect. For instance, we could observe the illumination on nose are significantly different.

7 LIMITATION

Although our method can accomplish a satisfactory makeup migration effect, we still found some noteworthy points in the process of experimentation, the makeup migration effect will be unevenly distributed when there is a large shadow area on the character's face, as shown in Figure 16. From the image, we can observe that there is a significant difference in the color depth between the left and right eyebrows of the subject, which diminishes the visual effectiveness of the makeup. We utilized the 3DMM to extract three-dimensional facial information, emphasizing its proficiency in capturing the geometry and surface details of the face. However, the model needs further enhancement in processing the complexities related to varying lighting conditions. In conditions of intense or uneven lighting, these complex features may be amplified or distorted, resulting in a misleading portrayal of makeup effects.

8 CONCLUSION

In this article, we propose a new makeup network called 2D and 3D Identity Preservation Net in an attempt to overcome several limitations of the current makeup transfer framework. IP23-Net first distills the facial geometric information and identity feature from the source image by 3D Shape Identity Encoder. Then, we leverage Makeup Style Encoder to extract the local makeup style from the reference image. Besides, a 3D Stereoscopic loss is proposed to provide structure supervision to achieve high-fidelity makeup transfer. Background Correction Decoder utilizes makeup and identity features for makeup transfer while restoring the background by distinguishing face and background using the face shape extracted through 3DMM. Moreover, we introduce a High Resolution Synthetic Makeup dataset by StyleGAN2 [32], which is the largest makeup dataset. Extensive experiments on the MT dataset demonstrate that our approach can achieve competitive makeup transfer results and preserve the background accurately.

REFERENCES

- [1] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 5799–5809. <https://doi.org/10.1109/CVPR46437.2021.00574>
- [2] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. 2018. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 40–48. <https://doi.org/10.1109/CVPR.2018.00012>

- [3] Cunjian Chen, Antitza Dantcheva, and Arun Ross. 2013. Automatic facial makeup detection with application in face recognition. In *Proceedings of the International Conference on Biometrics (ICB'13)*. IEEE, 1–8. <https://doi.org/10.1109/ICB.2013.6612994>
- [4] Cunjian Chen, Antitza Dantcheva, and Arun Ross. 2016. An ensemble of patch-based subspaces for makeup-robust face recognition. *Inf. Fusion* 32 (2016), 80–92. <https://doi.org/10.1016/J.INFFUS.2015.09.005>
- [5] Cunjian Chen, Antitza Dantcheva, Thomas Swearingen, and Arun Ross. 2017. Spoofing faces using makeup: An investigative study. In *Proceedings of the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA'17)*. IEEE, 1–8. <https://doi.org/10.1109/ISBA.2017.7947686>
- [6] Dong Chen, Yueting Zhuang, Zijin Shen, Carl Yang, Guoming Wang, Siliang Tang, and Yi Yang. 2022. Cross-modal data augmentation for tasks of different modalities. *IEEE Trans. Multimedia* (2022). <https://doi.org/10.1109/TMM.2022.3228696>
- [7] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. 2019. Beauty-glow: On-demand makeup transfer framework with reversible generative network. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 10042–10050. <https://doi.org/10.1109/CVPR.2019.01028>
- [8] Antitza Dantcheva, Cunjian Chen, and Arun Ross. 2012. Can facial cosmetics affect the matching accuracy of face recognition systems? In *Proceedings of the IEEE 5th International Conference on Biometrics: Theory, Applications and Systems (BTAS'12)*. IEEE, 391–398. <https://doi.org/10.1109/BTAS.2012.6374605>
- [9] Kirsten Dellinger and Christine L. Williams. 1997. Makeup at work: Negotiating appearance rules in the workplace. *Gender Soc.* 11, 2 (1997), 151–177. <https://doi.org/10.1177/089124397011002002>
- [10] Han Deng, Chu Han, Hongmin Cai, Guoqiang Han, and Shengfeng He. 2021. Spatially-invariant style-codes controlled makeup transfer. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 6549–6557. <https://doi.org/10.1109/CVPR46437.2021.00648>
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09)*. IEEE, 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [12] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'19)*. <https://doi.org/10.1109/CVPR.2019.00482>
- [13] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops'19)*. 0–0. <https://doi.org/10.1109/CVPRW.2019.00038>
- [14] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. 2021. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the International Conference on Computer Vision (ICCV'21)*. 14304–14313. <https://doi.org/10.1109/ICCV48922.2021.01404>
- [15] Yuhang Ding, Hehe Fan, Mingliang Xu, and Yi Yang. 2020. Adaptive exploration for unsupervised person re-identification. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 1 (2020), 1–19. <https://doi.org/10.1145/3369393>
- [16] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 4 (2018), 1–18. <https://doi.org/10.1145/3243316>
- [17] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. 2019. Ladrn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the International Conference on Computer Vision (ICCV'19)*. 10481–10490. <https://doi.org/10.1109/ICCV.2019.01058>
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [19] Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. 2020. Leveraging 2d data to learn textured 3d mesh generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 7498–7507. <https://doi.org/10.1109/CVPR42600.2020.00752>
- [20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS'17)*, Vol. 30. <https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fe65871369074926d-Abstract.html>
- [21] Daichi Horita and Kiyoharu Aizawa. 2022. SLGAN: Style-and latent-guided generative adversarial network for desirable makeup transfer and removal. In *Proceedings of the 4th ACM MM in Asia*. 1–5. <https://doi.org/10.1145/3551626.3564967>
- [22] Bingwen Hu, Zhedong Zheng, Ping Liu, Wankou Yang, and Mingwu Ren. 2020. Unsupervised eyeglasses removal in the wild. *IEEE Trans. Cybernet.* 51, 9 (2020), 4373–4385. <https://doi.org/10.1109/TCYB.2020.2995496>

- [23] Shuo Huang, Zongxin Yang, Liangting Li, Yi Yang, and Jia Jia. 2023. AvatarFusion: Zero-shot generation of clothing-decoupled 3D avatars using 2D diffusion. In *Proceedings of the ACM International Multimedia Conference (MM'23)*. 5734–5745. <https://doi.org/10.1145/3581783.3612022>
- [24] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the International Conference on Computer Vision (ICCV'17)*. 1501–1510. <https://doi.org/10.1109/ICCV.2017.167>
- [25] Zhikun Huang, Zhedong Zheng, Chenggang Yan, Hongtao Xie, Yaoqi Sun, Jianzhong Wang, and Jiyong Zhang. 2021. Real-world automatic makeup via identity preservation makeup net. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*. 652–658. <https://doi.org/10.24963/IJCAI.2020/91>
- [26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 1125–1134. <https://doi.org/10.1109/CVPR.2017.632>
- [27] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. 2020. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 5194–5202. <https://doi.org/10.1109/CVPR42600.2020.00524>
- [28] Wentao Jiang, Si Liu, Chen Gao, Ran He, Bo Li, and Shuicheng Yan. 2020. Beautify as you like. In *Proceedings of the 28th ACM International Multimedia Conference (MM'20)*. 4542–4544. <https://doi.org/10.1145/3394171.3414452>
- [29] Yang Jin, Wenhao Jiang, Yi Yang, and Yadong Mu. 2021. Zero-shot video event detection with high-order semantic concept discovery and matching. *IEEE Trans. Multimedia* 24 (2021), 1896–1908. <https://doi.org/10.1109/TMM.2021.3073624>
- [30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*. Springer, 694–711. https://doi.org/10.1007/978-3-319-46475-6_43
- [31] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 4401–4410. <https://doi.org/10.1109/CVPR.2019.00453>
- [32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 8110–8119. <https://doi.org/10.1109/CVPR42600.2020.00813>
- [33] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, Yoshua Bengio and Yann LeCun (Eds.).
- [34] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 5549–5558. <https://doi.org/10.1109/CVPR42600.2020.00559>
- [35] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. 2023. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the ACM International Multimedia Conference (MM'23)*. 1485–1494. <https://doi.org/10.1145/3581783.3611724>
- [36] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. 2018. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the ACM International Multimedia Conference (MM'18)*. 645–653. <https://doi.org/10.1145/3240508.3240618>
- [37] Xiangtai Li, Henghui Ding, Wenwei Zhang, Haobo Yuan, Guangliang Cheng, Pang Jiangmiao, Kai Chen, Ziwei Liu, and Chen Change Loy. 2023. Transformer-based visual segmentation: A survey (unpublished).
- [38] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. 2024. OMG-Seg: Is one model good enough for all segmentation? In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'24)*.
- [39] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual attribute transfer through deep image analogy. *ACM Trans. Graph.* 36, 4 (2017), 120:1–120:15. <https://doi.org/10.1145/3072959.3073683>
- [40] Si Liu, Wentao Jiang, Chen Gao, Ran He, Jiashi Feng, Bo Li, and Shuicheng Yan. 2021. Psgan++: Robust detail-preserving makeup transfer and removal. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 11 (2021), 8538–8551. <https://doi.org/10.1109/TPAMI.2021.3083484>
- [41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the International Conference on Computer Vision (ICCV'21)*. 10012–10022. <https://doi.org/10.1109/ICCV48922.2021.00986>
- [42] Yawei Luo and Yi Yang. 2024. Large language model and domain-specific model collaboration for smart education. *Front. Inf. Technol. Electr. Eng.* 1, 1 (2024). <https://doi.org/10.1631/FITEE.2300747>
- [43] Yueming Lyu, Jing Dong, Bo Peng, Wei Wang, and Tieniu Tan. 2021. SOGAN: 3D-aware shadow and occlusion robust GAN for makeup transfer. In *Proceedings of the 29th ACM International Multimedia Conference (MM)*. 3601–3609. <https://doi.org/10.1145/3474085.3475531>

- [44] Yueming Lyu, Yue Jiang, Ziwen He, Bo Peng, Yunfan Liu, and Jing Dong. 2023. 3D-aware adversarial makeup generation for facial privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2023), 13438–13453. DOI: <https://doi.org/10.1109/TPAMI.2023.3290175>
- [45] Thao Nguyen, Anh Tuan Tran, and Minh Hoai. 2021. Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 13305–13314. <https://doi.org/10.1109/CVPR46437.2021.01310>
- [46] Tam V Nguyen and Luoqi Liu. 2017. Smart mirror: Intelligent makeup recommendation and synthesis. In *Proceedings of the 25th ACM Proceedings of the ACM International Multimedia Conference (MM'17)*. 1253–1254. <https://doi.org/10.1145/3123266.3127926>
- [47] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. 2019. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the International Conference on Computer Vision (ICCV'19)*. 7588–7597. <https://doi.org/10.1109/ICCV.2019.00768>
- [48] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. 2020. Blockgan: Learning 3d object-aware scene representations from unlabelled images. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS'20)*, Vol. 33, 6767–6778.
- [49] Michael Niemeyer and Andreas Geiger. 2021. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 11453–11464. <https://doi.org/10.1109/CVPR46437.2021.01129>
- [50] Pietro Pala and Stefano Berretti. 2019. Reconstructing 3D face models by incremental aggregation and refinement of depth frames. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1 (2019), 1–24. <https://doi.org/10.1145/3287309>
- [51] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 296–301. <https://doi.org/10.1109/AVSS.2009.58>
- [52] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS'20)*, Vol. 33, 20154–20166.
- [53] Xiaolong Shen, Jianxin Ma, Chang Zhou, and Zongxin Yang. 2023. Controllable 3D face generation with conditional style code diffusion. arXiv:2312.13941. Retrieved from <https://arxiv.org/abs/2312.13941>
- [54] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2019. Increasing image memorability with neural style transfer. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 2 (2019), 1–22. <https://doi.org/10.1145/3311781>
- [55] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, Yoshua Bengio and Yann LeCun (Eds.).
- [56] Yucheng Suo, Zhedong Zheng, Xiaohan Wang, Bang Zhang, and Yi Yang. 2024. Jointly harnessing prior structures and temporal consistency for sign language video generation. *ACM Trans. Multimedia Comput. Commun. Appl.* (2024). <https://doi.org/10.1145/3648368>
- [57] Quan Wang, Sheng Li, Xinpeng Zhang, and Guorui Feng. 2022. Multi-granularity brushstrokes network for universal style transfer. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 4 (2022), 1–17. <https://doi.org/10.1145/3506710>
- [58] Xinrui Wang and Jinze Yu. 2020. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 8090–8099. <https://doi.org/10.1109/CVPR42600.2020.00811>
- [59] Xiaohan Wang, Linchao Zhu, Yu Wu, and Yi Yang. 2020. Symbiotic attention for egocentric action recognition with object-centric alignment. *IEEE Trans. Pattern Anal. Mach. Intell.* (2020). <https://doi.org/10.1109/TPAMI.2020.3015894>
- [60] Lin Xu, Yangzhou Du, and Yimin Zhang. 2013. An automatic framework for example-based virtual makeup. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 3206–3210. <https://doi.org/10.1109/ICIP.2013.6738660>
- [61] Chenyu Yang, Wanrong He, Yingqing Xu, and Yang Gao. 2022. Elegant: Exquisite and locally editable gan for makeup transfer. In *European Conference on Computer Vision*. Springer, 737–754. https://doi.org/10.1007/978-3-031-19787-1_42
- [62] Yi Yang, Yueting Zhuang, and Yunhe Pan. 2021. Multiple knowledge representation for big data artificial intelligence: Framework, applications, and case studies. *Front. Inf. Technol. Electr. Eng.* 22, 12 (2021), 1551–1558. <https://doi.org/10.1631/FITEE.2100463>
- [63] Zongxin Yang, Yunchao Wei, and Yi Yang. 2021. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 9 (2021), 4701–4712. <https://doi.org/10.1109/TPAMI.2021.3081597>
- [64] Lidong Zeng, Zhedong Zheng, Yinwei Wei, and Tat-seng Chua. 2024. Instilling multi-round thinking to text-guided image generation. arXiv:2401.08472. Retrieved from <https://arxiv.org/abs/2401.08472>

- [65] Kaihao Zhang, Dongxu Li, Wenhan Luo, Jingyu Liu, Jiankang Deng, Wei Liu, and Stefanos Zafeiriou. 2022. EDFace-celeb-1 M: Benchmarking face hallucination with a million-scale dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 3 (2022), 3968–3978. <https://doi.org/10.1109/TPAMI.2022.3181579>
- [66] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. 2022. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 18450–18459. <https://doi.org/10.1109/CVPR52688.2022.01790>
- [67] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Yi Yang, and Tat-Seng Chua. 2023. Multi-view consistent generative adversarial networks for compositional 3D-aware image synthesis. *Int. J. Comput. Vis.* (2023), 1–24.
- [68] Yi Zhang, Linzi Qu, Lihuo He, Wen Lu, and Xinbo Gao. 2019. Beauty aware network: An unsupervised method for makeup product retrieval. In *Proceedings of the 27th ACM International Multimedia Conference (MM'19)*. 2558–2562. <https://doi.org/10.1145/3343031.3356075>
- [69] Zhedong Zheng, Xiaohan Wang, Nenggan Zheng, and Yi Yang. 2022. Parameter-efficient person re-identification in the 3d space. *IEEE Trans. Neural Netw. Learn. Syst.* (2022). <https://doi.org/10.1109/TNNLS.2022.3214834>
- [70] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 1 (2017), 1–20. <https://doi.org/10.1145/3159171>
- [71] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. 2018. Camstyle: A novel data augmentation method for person re-identification. *IEEE Trans. Image Process.* 28, 3 (2018), 1176–1190. <https://doi.org/10.1109/TIP.2018.2874313>
- [72] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the International Conference on Computer Vision (ICCV'17)*. 2223–2232. <https://doi.org/10.1109/ICCV.2017.244>
- [73] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. 2018. Visual object networks: Image generation with disentangled 3D representations. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS'18)*, Vol. 31.

Received 21 August 2023; revised 20 February 2024; accepted 22 March 2024