Hybrid Transformers With Attention-Guided Spatial Embeddings for Makeup Transfer and Removal

Mingxiu Li, Wei Yu, Qinglin Liu, Zonglin Li, Ru Li, *Member, IEEE*,
Bineng Zhong, and Shengping Zhang, *Member, IEEE*

***Abstract*— Existing makeup transfer methods typically transfer simple makeup colors in a well-conditioned face image and fail to handle makeup style details (e.g., complicated colors and shapes) and facial occlusion. To address these problems, this paper proposes Hybrid Transformers with Attention-guided Spatial Embeddings (named HT-ASE) for makeup transfer and removal. Specifically, a makeup context extractor adopts makeup context global-local interactions to aggregate the high-level context and low-level detail features of the makeup styles, which obtains the context-aware makeup features that encode the complicated colors and shapes of the makeup styles. A face identity extractor adopts a face identity local interaction to aggregate the identity-relevant features of shallow layers into identity semantic features, which refines the identity features. A spatially similarity-aware fusion network introduces a spatially-adaptive layer-instance normalization with attention-guided spatial embeddings to perform semantic alignment and fusion between the makeup and identity features, yielding precise and robust transfer results even with large spatial misalignment and facial occlusion. Extensive experimental results demonstrate that the proposed method outperforms the state-of-the-art methods, especially in the preservation of makeup style details and handling facial occlusion.**

***Index Terms*— Makeup transfer, makeup removal, vision transformer.**

I. Introduction

MAKEUP transfer has been extensively studied in computer vision [1], [2], [3], [4], [5], [6] and graphics [7], [8], [9] since it has broad demands in many popular beauty applications such as TikTok [10] and MeiTu [11] Moreover, it provides inspiration for exciting research on face verification and attack [12], [13]. The goal of makeup transfer is to render a source face with a reference makeup style while preserving the original face identity. To this end, four challenging issues must be considered: (1) To extract the makeup styles with complicated details from the reference face image, including the color and shape of eye shadows, the highlight on nose, the blush on cheeks, etc. (2) To preserve the identity features of the source face image, such as facial structure, wrinkles around eyes, and skin texture details. (3) To perform customized makeup transfer, such as the flexibility of intensity controllable and partial makeup transfer. (4) To be robust to facial occlusion and spatial misalignments, such as a hand on face and different expressions/poses between two faces.

Traditional methods adopt physics-based reflectance models [8], [9] and image gradient editing [7] for makeup transfer. Recently, deep learning methods [1], [2], [3], [5], [6], [14], [15] based on generative adversarial networks [16] and disentangled representation networks [17] make significant progress in transferring relatively simple makeup colors to facial regions in a well-conditioned clean face image. However, these methods are still hampered by two practical problems. First, they fail to transfer makeup styles with complicated details from the reference face image, including the color and shape of eye shadows, lipstick color, blush on cheeks, and highlights on noses, as shown in Fig. 1a. In addition, when a facial occlusion exists in the reference face image, these methods usually fail to handle the details of the makeup styles and can not repair the occluded regions, which causes severe artifacts in the generated images. As shown in Fig. 1b, the eye shadows and skin generated by the state-of-the-art methods are unnatural.

The reasons why the above problems happen may be threefold: (1) Although feature maps at multiple scales contain different contextual and semantic information, existing methods do not design reasonable feature interaction modules for the extracted multi-scale features, which leads to a large amount of contextual and semantic information lost in the extracted makeup features and identity features, especially for the complicated colors and shapes of the makeup styles. (2) Existing methods fail to establish an accurate semantic correspondence between the two faces, resulting in makeup details being transferred to wrong semantic positions and the inability to repair the occluded facial regions. In particular, they fail to preserve the texture and semantic information of the identity features that can complement the semantic matching evidence and improve the performance of semantic similarity learning between the two face images. (3) Existing methods neglect to maintain the spatial and semantic information of the makeup features when fusing the makeup features

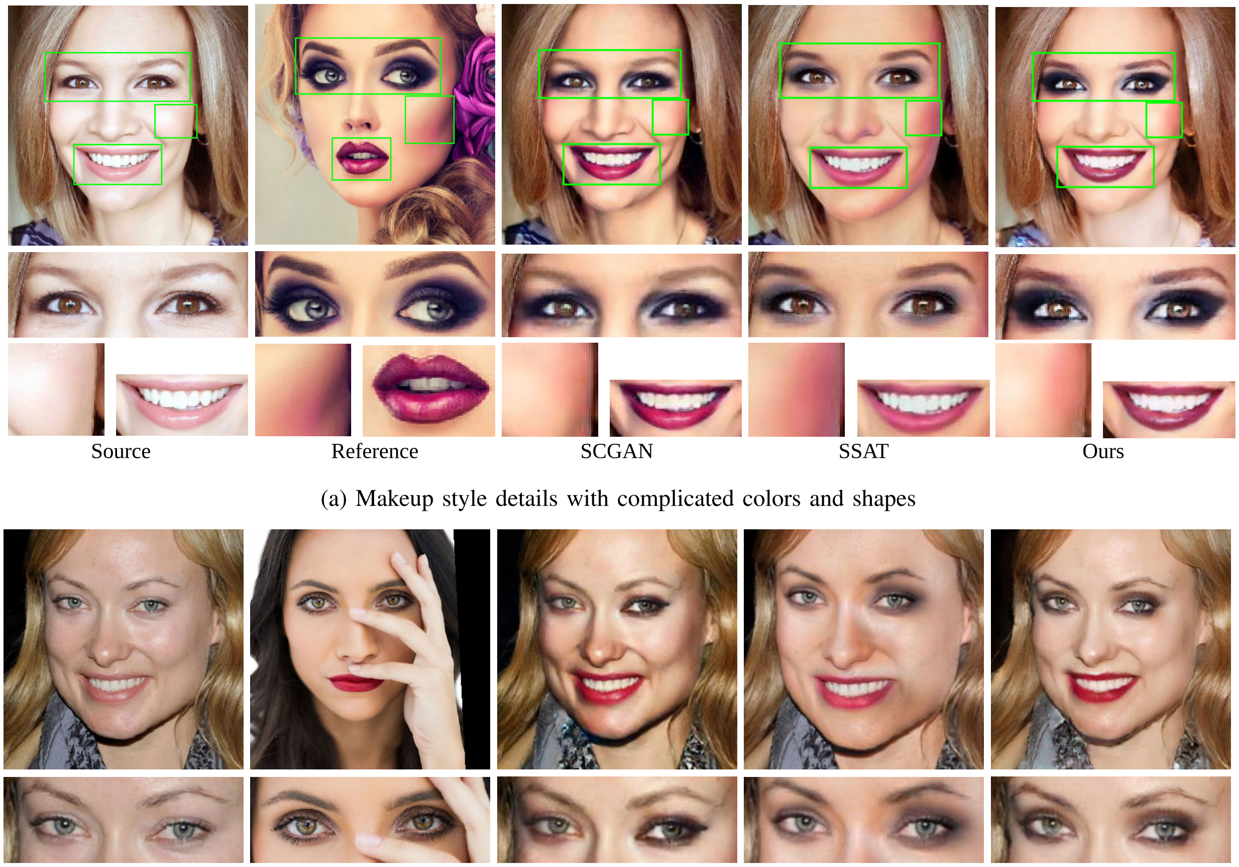
Manuscript received 12 April 2023; revised 11 August 2023; accepted 2 September 2023. Date of publication 11 September 2023; date of current version 5 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62272134 and in part by the Taishan Scholars Program of Shandong Province under Grant tsqn201812106. This article was recommended by Associate Editor Z. Qian. *(Corresponding author: Qinglin Liu.)*

Mingxiu Li, Wei Yu, Qinglin Liu, Zonglin Li, Ru Li, and Shengping Zhang are with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, China (e-mail: 20B903058@stu.hit.edu.cn; 20B903014@stu.hit.edu.cn; qinglin.liu@outlook.com; zonglin.li@hit.edu.cn; liru@hit.edu.cn; s.zhang@hit.edu.cn).

Bineng Zhong is with the School of Computer Science and Engineering, Guangxi Normal University, Guilin 541004, China (e-mail: bnzhong@gxnu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3312790.

Digital Object Identifier 10.1109/TCSVT.2023.3312790



(a) Makeup style details with complicated colors and shapes

(b) Makeup styles with facial occlusion

Fig. 1. Examples of the makeup transfer results of the proposed HT-ASE and two state-of-the-art methods including SCGAN [1], and SSAT [2].

and identity features, which leads to the failure of preserving the shapes of the makeup details in the transferred results.

To address the above-mentioned problems, this paper proposes Hybrid Transformers with Attention-guided Spatial Embeddings (named HT-ASE) to extract makeup styles and face identities through multi-scale feature interactions and establish accurate semantic correspondences between the source and reference faces to generate precise and robust makeup transfer results. Specifically, a makeup context extractor adopts makeup context global-local interactions to aggregate the high-level context and low-level detail features of the makeup styles, which obtains the context-aware makeup features that encode the complicated colors and shapes of the makeup styles. A face identity extractor adopts a face identity local interaction to aggregate the identity-relevant features of shallow layers into identity semantic features, which refines the identity features. A spatially similarity-aware fusion network introduces a spatially-adaptive layer-instance normalization with attention-guided spatial embeddings to perform semantic alignment and fusion between the makeup and identity features, yielding precise and robust transfer results even with large spatial misalignment and facial occlusion. Extensive qualitative and quantitative experiments demonstrate that our proposed method outperforms the state-of-the-art methods for makeup transfer and removal, especially in the preservation of makeup style details and handling facial occlusion.

The main contributions of this paper can be summarized as follows:

- We propose Hybrid Transformers with Attention-guided Spatial Embeddings (named HT-ASE) to perform multi-scale feature interactions and establish semantic correspondences between the source and reference face images, which generates precise and robust makeup transfer and removal results.
- We design a makeup context extractor with makeup context global-local interactions to aggregate the high-level context and low-level detail features of makeup styles, which learns context-aware makeup features that encode complicated colors and shapes of makeup styles.
- We design a spatially-adaptive layer-instance normalization with attention-guided spatial embeddings to perform semantic alignment and fusion between the makeup and identity features even with large spatial misalignment and facial occlusion.
- Extensive qualitative and quantitative experiments demonstrate that the proposed method outperforms the leading methods, especially in the preservation of makeup style details and handling facial occlusion.

II. Related Work

A. Makeup Transfer and Removal

Makeup transfer aims to render a source face with a reference makeup style while preserving the original face identity, which has been extensively studied [1], [2], [3], [5], [6], [14], [15], [18], [19], [20], [21]. BeautyGAN [18] proposes a dual-input/output GAN [16] to simultaneously achieve makeup transfer and removal. Besides, it introduces the pixel-level histogram loss to calculate each makeup region separately, followed by most makeup transfer approaches. LADN [14] introduces multiple overlapping local discriminators and asymmetric losses to achieve dramatic makeup transfer and removal. The above methods only perform makeup transfer on frontal-aligned faces with neutral expressions. To transfer makeup styles between images with large spatial misalignment, PSGAN [15], PSGAN++ [6] and FAT [5] use the facial landmarks and attention mechanism to establish pixel-level correspondences for identity features. SCGAN [1] uses a non-linear mapping network to remove the spatial information of makeup features, which solves the problem of spatial misalignment between images and achieves controllable makeup styles. SSAT [2] extracts semantic information through binary face parsing maps, which facilitates transferring makeup styles to more appropriate semantic locations. The makeup transfer methods mentioned above typically transfer simple makeup colors in a well-conditioned clean image and fail to handle makeup style details (e.g., complicated colors and shapes) and facial occlusion. In contrast, we perform cross-layer feature interactions and establish semantic correspondences between the source and reference faces to generate precise and robust makeup transfer and removal results.

B. Style Transfer

Style transfer can be considered as a general form of makeup transfer. Recently, style transfer methods based on deep convolutional neural networks have been extensively studied and made great progress [22], [23], [24], [25], [26], [27], [28]. Reference [22] proposes a framework based on multi-label semantics and GAN for perception-driven wallpaper texture generation and style transfer. Reference [28] proposes a neural distortion field and a neural texture transformation network to distort the source shape to the geometric style of the target and transfer the artistic style to the distorted source product, respectively. For the quality assessment of arbitrary neural style transfer (AST) images, [26] presents a sparse representation-based method, which calculates the quality according to the similarity of sparse features. Style transfer methods usually require the ability to transfer the style from one domain to another. However, the lack of the ability to control and understand the local details in the styles and contents makes it unsuitable for application in makeup transfer and removal. In contrast, we design a makeup context extractor with makeup context global-local interactions to aggregate the high-level context and low-level detail features of makeup styles, which learns context-aware makeup features that encode the complicated colors and shapes of makeup styles.

C. Vision Transformer Models

Transformer focuses on important image regions by modeling long-range dependencies, which has gained a lot of attention and shown good performance in some vision tasks [29], [30], [31], [32], [33]. Transformer is introduced into image style transfer. StyTr2 [31] is based on a transformer architecture that takes the long-distance dependency of the input image into account for style transfer. StyleFormer [32] is a feed-forward multiple style transfer method with a transformer architecture that guarantees both semantic content consistency and fine-grained style diversity. Swin Transformer [33] shows excellent promise by integrating the advantages of CNN and transformer, which uses a local attention mechanism to process large-scale images and establishes long-range dependencies through a shifted window scheme. Therefore, we introduce the structure of the transformer for makeup transfer and removal.

III. Our Approach

A. Task Definition

Makeup transfer aims to render a source face image with a reference makeup style while preserving the original face identity. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the non-makeup and makeup image domains, respectively. Let $\{\mathbf{X}_n\}_{n=1}^{N}$ and $\{\mathbf{Y}_m\}_{m=1}^{M}$ be the sampled images from $\mathcal{X}$ and $\mathcal{Y}$, respectively. We assume the makeup images have different face identities with the source face images. Given the source face image $\mathbf{X}_n$ and reference face image $\mathbf{Y}_m$, makeup transfer aims to learn a mapping function: $\hat{\mathbf{X}}_n = \mathrm{G}(\mathbf{X}_n, \mathbf{Y}_m)$, where $\hat{\mathbf{X}}_n$ has the face identity of $\mathbf{X}_n$ and the makeup style of $\mathbf{Y}_m$. Makeup removal is assumed to be a special case of makeup transfer [34], which considers non-makeup images with special makeup styles and transfers these special makeup styles to makeup images. Therefore, makeup removal aims to learn a mapping function: $\hat{\mathbf{Y}}_m = \mathrm{G}(\mathbf{Y}_m, \mathbf{X}_n)$, where $\hat{\mathbf{Y}}_m$ has the facial identity of $\mathbf{Y}_m$ and the makeup style of $\mathbf{X}_n$.

B. Overall Architecture

Fig. 2 shows the overall architecture of the proposed HT-ASE, which contains three sub-networks: a makeup context extractor, a face identity extractor, and a spatially similarity-aware fusion network. Given the source face image $\mathbf{X}_n$ and reference face image $\mathbf{Y}_m$, the makeup context extractor adopts a makeup feature pyramidal encoder and a makeup context global-local interaction module to extract makeup features $\mathbf{X}_n^s$ and $\mathbf{Y}_m^s$ from two face images, respectively. The face identity extractor adopts an identity feature pyramidal encoder and a face identity local interaction module to extract identity features $\mathbf{X}_n^c$ and $\mathbf{Y}_m^c$ from two face images, respectively. Then, the spatially similarity-aware fusion network adaptively fuses the makeup feature $\mathbf{Y}_m^s$ and identity feature $\mathbf{X}_n^c$ under the correlation constraints of the identity features $\mathbf{X}_n^c$ and $\mathbf{Y}_m^c$ to obtain the makeup transfer result $\hat{\mathbf{X}}_n$. We use the same procedure to obtain the makeup removal result $\hat{\mathbf{Y}}_m$.

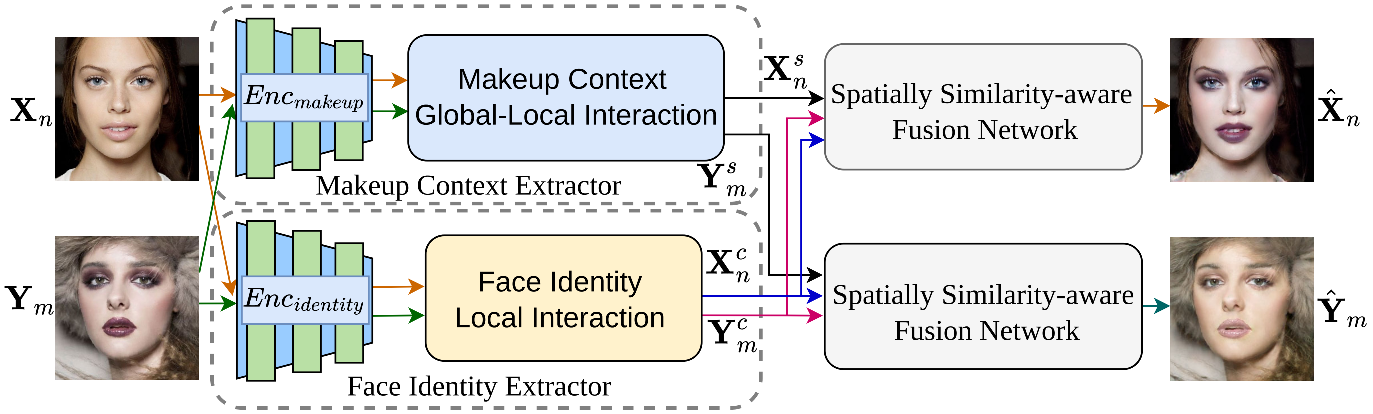


Fig. 2. Overview framework of the proposed HT-ASE for makeup transfer and removal. The process is executed in the following steps: 1) The makeup context encoder extracts the makeup features of the source face image $\mathbf{X}_n$ and reference face image $\mathbf{Y}_m$. 2) The face identity encoder extracts identity features of the source face image $\mathbf{X}_n$ and reference face image $\mathbf{Y}_m$. 3) The spatially similarity-aware fusion network adaptively fuses identity features $\mathbf{X}_n^c, \mathbf{Y}_m^c$ with makeup features $\mathbf{X}_n^s, \mathbf{Y}_m^s$ to generate the makeup transfer result $\hat{\mathbf{X}}_n$ and the makeup removal result $\hat{\mathbf{Y}}_m$.

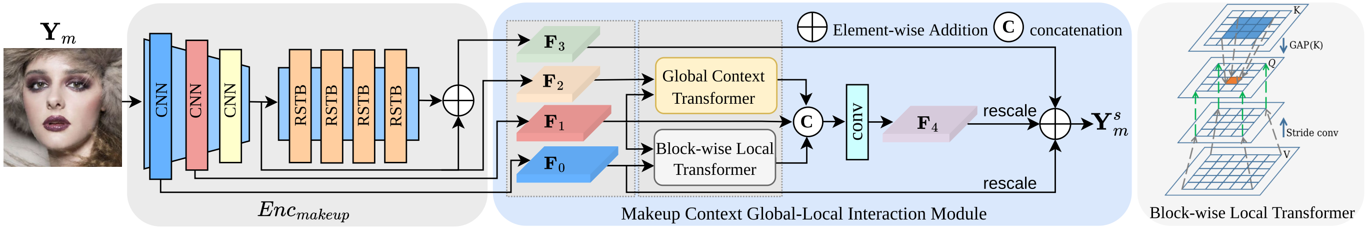


Fig. 3. Detailed architecture of the proposed makeup context extractor, which includes a makeup feature pyramidal encoder Enc_{makeup} and a makeup context global-local interaction module to obtain makeup features that encode the complicated colors and shapes of makeup styles.

C. Makeup Context Extractor

Since existing makeup transfer methods typically transfer simple makeup colors in a well-conditioned face image and fail to handle makeup style details (e.g., complicated colors and shapes), we design a makeup context extractor to address these problems. As shown in Fig. 3, the core components of the makeup context extractor are a makeup feature pyramidal encoder and a makeup context global-local interaction module. The hybrid architecture based on CNN and transformers aggregates the low-level detail features and the high-level context features of makeup styles to obtain context-aware makeup features that encode the complicated colors and shapes of makeup styles.

Since there is domain deviation between the makeup features and identity features, the feature pyramidal encoders with three downsampling convolution layers and five residual swin transformer blocks (RSTB) [35] only extract a few domain-specific features, which is not suitable enough for makeup transfer and removal [36]. To strengthen the disentanglement ability, the makeup context extractor designs a makeup context global-local interaction module to incorporate multi-level contextual information. Specifically, a global context transformer (GCT) and a block-wise local transformer (BLT) are introduced into the makeup context global-local interaction module to implement global context-guided local makeup extraction. The GCT integrates $\mathbf{F}_2$ to $\mathbf{F}_1$ and the BLT uses $\mathbf{F}_0$ to render $\mathbf{F}_1$. The output features of two transformers and the original feature $\mathbf{F}_1$ from the makeup feature pyramidal encoder are concatenated and go through a convolution operation to obtain the complementary feature $\mathbf{F}_4$. Considering that $\mathbf{F}_0$ has the largest spatial resolution and contains rich information on high-frequency makeup details, $\mathbf{F}_4$ has rich contextual and semantic information through cross-layer global-local feature interactions, they are rescaled and added with $\mathbf{F}_4$ to obtain the makeup feature $\mathbf{Y}_m^s$.

In particular, the GCT and BLT interact the shallow features $\mathbf{F}_0$, $\mathbf{F}_1$ and $\mathbf{F}_2$ to aggregate the makeup-relevant features extracted from adjacent convolution layers, which better encode complicated colors and shapes of makeup styles. The GCT renders the global contextual and semantic information of higher-level feature map $\mathbf{F}_2$ into the feature points of lower-level feature map $\mathbf{F}_1$, which implements global context-guided makeup styles extraction. GCT is formulated as

$$f(\mathbf{F}_1, \mathbf{F}_2) = \frac{1}{N}(\mathbf{W}_q\mathbf{F}_1)^T(\mathbf{W}_k\mathbf{F}_2)$$
$$\hat{\mathbf{F}}_1 = f(\mathbf{F}_1, \mathbf{F}_2)(\mathbf{W}_v\mathbf{F}_2) \tag{1}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are the parameters of query, key, and value, respectively. N is an invariant constant whose value is the size of the flattened attentive matrix calculated by the query and key. The BLT uses the local high-frequency details of the lower-level feature map $\mathbf{F}_0$ to enrich the high-frequency information of the corresponding pixels in $\mathbf{F}_1$. First, the higher-level feature map $\mathbf{F}_1$ is transformed to $\mathbf{Q} = \mathrm{H}_q(\mathbf{F}_1)$. The lower-level feature map $\mathbf{F}_0$ is transformed to $\mathbf{K} = \mathrm{H}_k(\mathbf{F}_0)$ and $\mathbf{V} = \mathrm{H}_v(\mathbf{F}_0)$, where $\mathrm{H}_q(\cdot)$, $\mathrm{H}_k(\cdot)$, and $\mathrm{H}_v(\cdot)$ denote three transformations that convolution layers can easily implement. Then, we use $\mathbf{K}$ and a Global Average Pooling (GAP) to

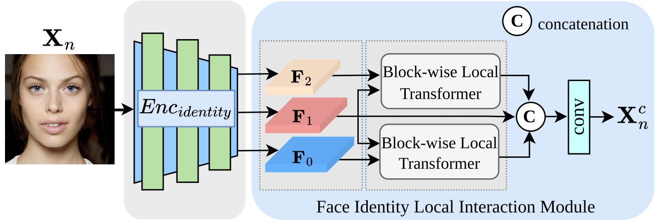


Fig. 4. Detailed architecture of the proposed face identity extractor, which includes an identity feature pyramidal encoder $Enc_{identity}$ and a face identity local interaction module to obtain the identity feature with high-frequency identity details and semantic identity information.

calculate a channel attention weight [37] for $\mathbf{Q}$. Finally, $\mathbf{V}$ is reduced the feature scale by one convolution layer and added to the weighted $\mathbf{Q}$. BLT is formulated as

$$X = \mathrm{SCONV}(\mathbf{V}) + \mathrm{GAP}(\mathbf{K}) \odot \mathbf{Q} \tag{2}$$

where $\mathrm{SCONV}(\cdot)$ is a 3×3 convolution layer with a 2×2 stride and $\odot$ indicates the Hadamard product.

D. Face Identity Extractor

As shown in Fig. 4, the face identity extractor adopts a hybrid architecture based on CNN and transformers, which encodes and integrates the high-frequency details of low-level face identity features extracted from shallow convolution layers into the high-level semantic feature extracted from RSTB. In particular, two BLTs are introduced into the face identity local interaction module, which use the local high-frequency details in shallow features $\mathbf{I}_0$ and $\mathbf{I}_1$ to enrich the high-frequency information of the corresponding pixels in $\mathbf{I}_2$. Then, the output features of BLTs and the original feature $\mathbf{I}_2$ are concatenated and go through a convolution layer to obtain the identity feature with high-frequency identity details and semantic identity information. The face identity feature $\mathbf{X}_n^c$ of the source face image $\mathbf{X}_n$ and the face identity feature $\mathbf{Y}_m^c$ of the reference face image $\mathbf{Y}_m$ can be respectively extracted by the face identity extractor.

E. Spatially Similarity-Aware Fusion Network

With the extracted makeup features and identity features from two extractors, a spatially similarity-aware fusion network (SSFNet) introduces a spatially-adaptive layer-instance normalization with attention-guided spatial embeddings to perform semantic alignment between the makeup and identity features under the correlation constraints of the identity features and then adaptively fuse the aligned makeup features and identity features to generate the transferred results even with large spatial misalignment and facial occlusion. As shown in Fig. 5 (a), the core components of SSFNet contain a semantic similarity learning (SSL) module and multiple residual adaptive fusion blocks with spatially-adaptive layer-instance normalization (SAdaLIN).

1) Semantic Similarity Learning (SSL): As shown in Fig. 5(b), a semantic similarity learning (SSL) module is introduced to learn a semantic similarity matrix $\mathbf{S} \in \mathbb{R}^{hw\times hw}$ of two identity features $\mathbf{X}_n^c$ and $\mathbf{Y}_m^c$, which helps to model the spatial position mapping of the two faces and spatially warp the makeup details in the feature $\mathbf{X}_n^s$ to the accurate semantic positions of the source face image $\mathbf{X}_n$ while preserving the shapes of makeup details. The design of SSL helps our proposed method robustly handle images with large spatial misalignment and facial occlusion. First, we calculate the visual correlation of each point in the source face identity feature $\mathbf{X}_n^c$ with all points in the reference face identity feature $\mathbf{Y}_m^c$. We reshape $\mathbf{X}_n^c$ and $\mathbf{Y}_m^c$ into $\hat{\mathbf{X}}_n^c = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{HW}] \in \mathbb{R}^{C\times HW}$ and $\hat{\mathbf{Y}}_m^c = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{HW}] \in \mathbb{R}^{C\times HW}$, respectively, where $\mathbf{a}_u \in \mathbb{R}^C$ indicates the feature vector of the u-th location in the source face image $\mathbf{X}$, $\mathbf{b}_e \in \mathbb{R}^C$ indicates the feature vector of the e-th location in the reference face image $\mathbf{Y}$. To enhance the stability of learning [38], the channel-wise features $\mathbf{a}_u$ and $\mathbf{b}_e$ are centralized to obtain $\hat{\mathbf{a}}_u$ and $\hat{\mathbf{b}}_e$ by subtracting the mean value of their respective feature maps. Furthermore, we use the method in [39] to obtain the face parsing regions of the two face images, which efficiently supervises the learning of semantic similarity matrix $\mathbf{S}$ through the divided facial regions (eyes, skin, and lips). The value $\mathbf{S}_{u,e}$ of the semantic similarity matrix $\mathbf{S} \in \mathbb{R}^{hw\times hw}$ is calculated by

$$\mathbf{S}_{u,e} = \frac{(\frac{\hat{\mathbf{a}}_u^T \hat{\mathbf{b}}_e}{\|\hat{\mathbf{a}}_u\|\|\hat{\mathbf{b}}_e\|})\mathbb{I}(\mathbf{R}^u_{\mathbf{X}_n^c} = \mathbf{R}^e_{\mathbf{Y}_m^c})}{\sum_{u=0}^{U-1}(\frac{\hat{\mathbf{a}}_u^T \hat{\mathbf{b}}_e}{\|\hat{\mathbf{a}}_u\|\|\hat{\mathbf{b}}_e\|})\mathbb{I}(\mathbf{R}^u_{\mathbf{X}_n^c} = \mathbf{R}^e_{\mathbf{Y}_m^c})} \tag{3}$$

where $\mathbb{I}(\cdot)$ is an indicator function whose value is 1 if the inside formula is true, $\mathbf{R}^u_{\mathbf{X}_n^c}$ and $\mathbf{R}^e_{\mathbf{Y}_m^c}$ denote the facial regions (eyes, skin and lips) that the u-th and e-th positions in $\mathbf{X}_n^c$ and $\mathbf{Y}_m^c$ belong to.

Fig. 6 shows the learned semantic similarity maps for a particular point, which illustrates pixel-wise contextual attention for this point on the source face image with all points on the reference face image. The first image shows the source face image and an example pixel, i.e., the red point on the upper right corner of the lip. The middle image shows the learned semantic similarity map for this pixel before applying softmax, which is obtained through reshaping a specific row of the semantic similarity matrix $\mathbf{S}$ with the shape $H \times W$. Note that there are many large attention weights around the lips. The last image shows the learned semantic similarity map for this pixel after applying softmax, where the attention weights are more concentrated on the lower edge of the lips. The above experiments demonstrate the ability of SSL to locate and focus on the points associated with a particular point.

2) Spatially-Adaptive Layer-Instance Normalization: Inspired by the previous work that incorporates different normalization functions and uses affine transformation parameters in normalization layers [40], [41], we design a spatially-adaptive layer-instance normalization (SAdaLIN) with attention-guided spatial embeddings to perform semantic alignment between the makeup and identity features and then adaptively fuse the aligned makeup features and identity features even with large spatial misalignment and facial occlusion. As shown in Fig. 5 (a), each residual adaptive fusion block contains two SAdaLINs. Fig. 5 (c) shows the detailed design of the proposed SAdaLIN, we consider to integrate the advantages of layer normalization (LN) [42] and adaptive instance

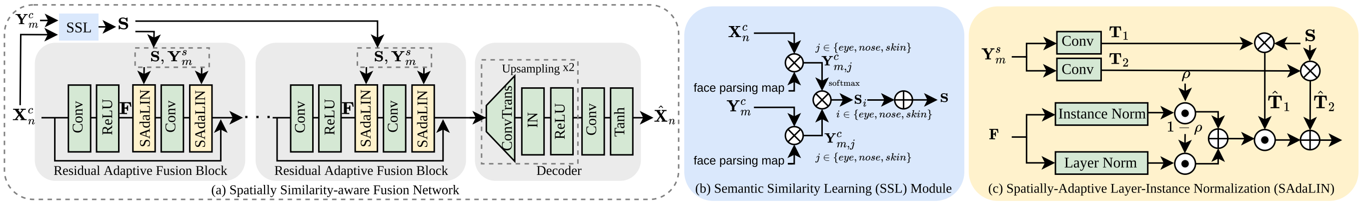


Fig. 5. Illustration of spatially similarity-aware fusion network. First, a semantic similarity learning (SSL) module learns a semantic similarity matrix $\mathbf{S} \in \mathbb{R}^{hw\times hw}$ between the identity features $\mathbf{X}_n^c$ and $\mathbf{Y}_m^c$ to model spatial position mapping of the two faces. Then, multiple residual adaptive fusion blocks with spatially-adaptive layer-instance normalization (SAdaLIN) adaptively fuse makeup feature $\mathbf{Y}_m^s$ with face identity feature $\mathbf{X}_n^c$ to generate a transfer result, even with large spatial misalignment and facial occlusion.

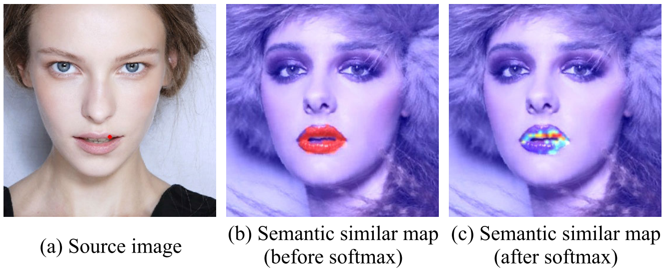

(a) Source image
(b) Semantic similar map (before softmax)
(c) Semantic similar map (after softmax)

Fig. 6. The attention maps for the red point on the source face image. Guided by the face resolution maps, the attention weights between the specific red point and pixels in the same face region are computed.

normalization (AdaIN) [43] into SAdaLIN, which enables SAdaLIN to keep or change the content optionally. On the one hand, LN considers the global statistics of feature maps. On the other hand, AdaIN preserves the content structure of the source domain and assumes no correlation between channels. Since AdaIN represents makeup features as 1D-vectors through fully connected layers, which results in the spatial information of makeup styles being lost and is detrimental to the preservation of high-frequency makeup details, such as the shape of eye shadow, we map the makeup feature $\mathbf{Y}_m^s \in \mathbb{R}^{C\times H\times W}$ to produce two makeup features $\mathbf{T}_1 \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{T}_2 \in \mathbb{R}^{C\times H\times W}$ as the affine transformation parameters of makeup styles through 1×1 convolution layers. $\mathbf{T}_1$ and $\mathbf{T}_2$ are respectively multiplied with the semantic similarity matrix $\mathbf{S}$, which introduces attention-guided spatical embeddings and makes $\mathbf{T}_1$ and $\mathbf{T}_2$ semantically aligned with the source face image $\mathbf{X}_n$. The morphed makeup features $\hat{\mathbf{T}}_1 \in \mathbb{R}^{C\times H\times W}$ and $\hat{\mathbf{T}}_2 \in \mathbb{R}^{C\times H\times W}$ are computed by

$$\hat{\mathbf{T}}_1^u = \sum_e \text{softmax}(\mathbf{S}_{u,e}) \cdot \mathbf{T}_1^e$$
$$\hat{\mathbf{T}}_2^u = \sum_e \text{softmax}(\mathbf{S}_{u,e}) \cdot \mathbf{T}_2^e \tag{4}$$

where u and e indicate the location index of $\mathbf{X}_n$ and $\mathbf{Y}_m$, respectively. Our proposed normalization technique SAdaLIN is defined as

$$\text{SAdaLIN}\left(\rho, \hat{\mathbf{T}}_1, \hat{\mathbf{T}}_2\right) = \hat{\mathbf{T}}_1 \cdot \left(\rho \cdot \hat{\mathbf{F}}_\text{I} + (1-\rho) \cdot \hat{\mathbf{F}}_\text{L}\right) + \hat{\mathbf{T}}_2 \tag{5}$$

After using multiple residual adaptive fusion blocks with SAdaLINs to fuse makeup features and identity features, we use a decoder with two upsampling convolution layers to generate the makeup transfer result $\hat{\mathbf{X}}_n$ and the makeup removal result $\hat{\mathbf{Y}}_m$, respectively.

F. Objective Function

1) Adversarial Loss: Two discriminators $\text{D}_\mathcal{X}$ and $\text{D}_\mathcal{Y}$ are introduced to discriminate between the generated images and real images in the non-makeup image domain $\mathcal{X}$ and the makeup image domain $\mathcal{Y}$. We add an attention mechanism to the discriminators, which has the same structure as [41]. The adversarial losses [44] $\mathcal{L}_\text{G}^{adv}$ and $\mathcal{L}_\text{D}^{adv}$ for the generator and discriminators are defined as

$$\begin{aligned}\mathcal{L}_\text{D}^{adv} &= -\mathbb{E}_{\mathbf{X}\sim\mathcal{X}}[\log\text{D}_\mathcal{X}(\mathbf{X})] - \mathbb{E}_{\mathbf{Y}\sim\mathcal{Y}}[\log\text{D}_\mathcal{Y}(\mathbf{Y})]\\ &\quad - \mathbb{E}_{\mathbf{X}\sim\mathcal{X},\mathbf{Y}\sim\mathcal{Y}}[\log(1-\text{D}_\mathcal{Y}(\text{G}(\mathbf{X},\mathbf{Y}))]\\ &\quad - \mathbb{E}_{\mathbf{X}\sim\mathcal{X},\mathbf{Y}\sim\mathcal{Y}}[\log(1-\text{D}_\mathcal{X}(\text{G}(\mathbf{Y},\mathbf{X}))]\\ \mathcal{L}_\text{G}^{adv} &= -\mathbb{E}_{\mathbf{X}\sim\mathcal{X},\mathbf{Y}\sim\mathcal{Y}}[log(\text{D}_\mathcal{Y}(\text{G}(\mathbf{X},\mathbf{Y})))]\\ &\quad - \mathbb{E}_{\mathbf{X}\sim\mathcal{X},\mathbf{Y}\sim\mathcal{Y}}[log(\text{D}_\mathcal{X}(\text{G}(\mathbf{Y},\mathbf{X})))]\end{aligned} \tag{6}$$

2) Cycle Consistency Loss: Due to the lack of paired makeup images and non-makeup images, the training process of the proposed HT-ASE follows CycleGAN [45] to learn the mapping between two domains bidirectionally. The cycle consistency loss is defined as

$$\mathcal{L}_\text{G}^{cyc} = \|\text{G}(\text{G}(\mathbf{X},\mathbf{Y}),\mathbf{X}) - \mathbf{X}\|_1 + \|\text{G}(\text{G}(\mathbf{Y},\mathbf{X}),\mathbf{Y}) - \mathbf{Y}\|_1 \tag{7}$$

where $\|\cdot\|_1$ denotes the L_1-Norm.

3) Perceptual Loss: A perceptual loss [46] is introduced to compare the discrepancies between two images in deep layers to ensure the consistency of facial identities between the generated face image and source face image. The perceptual loss is defined as

$$\mathcal{L}_\text{G}^{per} = \|\text{F}_l(\text{G}(\mathbf{X},\mathbf{Y})) - \text{F}_l(\mathbf{X})\|_2 + \|\text{F}_l(\text{G}(\mathbf{Y},\mathbf{X})) - \text{F}_l(\mathbf{Y})\|_2 \tag{8}$$

where $\|\cdot\|_2$ denotes the L_2-Norm and $\text{F}_l(\cdot)$ denotes the output of the l-th layer in the VGG model [47].

4) Local Invariance Loss: Since the human perceptual system is very sensitive to facial artifacts, we use L_1 loss to constrain eyeballs and teeth. The local invariance loss is defined as

$$\begin{aligned}\mathcal{L}_\text{G}^{local} &= \|(\text{G}(\mathbf{X},\mathbf{Y}) \odot \mathbf{R}_\mathbf{X} - \mathbf{X} \odot \mathbf{R}_\mathbf{X}\|_1\\ &\quad + \|(\text{G}(\mathbf{Y},\mathbf{X}) \odot \mathbf{R}_\mathbf{Y} - \mathbf{Y} \odot \mathbf{R}_\mathbf{Y}\|_1\end{aligned} \tag{9}$$

TABLE I

COMPARISON OF HT-ASE AND STATE-OF-THE-ART METHODS

Method	Capability		Controllability	
	Misalignent	Detail	Shade	Part
BeautyGAN [18]				
LADN [14]			✓	
PSGAN [15]	✓		✓	✓
SCGAN [1]	✓		✓	✓
SSAT [2]	✓			✓
HT-ASE (ours)	✓	✓	✓	✓

where $\odot$ denotes the Hadamard product, $\mathbf{R_X}$ and $\mathbf{R_Y}$ denote the regions of teeth and eyeballs in the source and reference face images, respectively.

5) Makeup Loss: To guide makeup transfer, we adopt the makeup loss [18], which applies Histogram Matching (HM) on each of the three parts: skin, lips, and eyes. Then we integrate them into a pseudo ground truth HM$(\cdot, \cdot)$. The makeup loss is defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{G}}^{makeup} &= \|\mathrm{G}(\mathbf{X}, \mathbf{Y}) - \mathrm{HM}(\mathbf{X}, \mathbf{Y})\|_2 \\ &\quad + \|\mathrm{G}(\mathbf{Y}, \mathbf{X}) - \mathrm{HM}(\mathbf{Y}, \mathbf{X})\|_2 \end{aligned} \tag{10}$$

6) Total Loss: The total loss L_D and L_G for the discriminator and generator of our HT-ASE are defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{D}} &= \lambda_{adv}\mathcal{L}_{\mathrm{D}}^{adv} \\ \mathcal{L}_{\mathrm{G}} &= \lambda_{adv}\mathcal{L}_{\mathrm{G}}^{adv} + \lambda_{cyc}\mathcal{L}_{\mathrm{G}}^{cyc} + \lambda_{per}\mathcal{L}_{\mathrm{G}}^{per} \\ &\quad + \lambda_{local}\mathcal{L}_{\mathrm{G}}^{local} + \lambda_{makeup}\mathcal{L}_{\mathrm{G}}^{makeup} \end{aligned} \tag{11}$$

where $\lambda_{adv}, \lambda_{cyc}, \lambda_{per}, \lambda_{local}$, and λ_{makeup} are the weights of the loss terms, respectively.

IV. EXPERIMENTS

A. Experimental Setting

We train and test the proposed HT-ASE on the Makeup Transfer (MT) dataset [18], which contains 1,115 non-makeup images and 2,719 makeup images. These 3,834 female images with the resolution of 361 × 361 contain various expressions, poses, light to heavy makeup styles, etc. We follow the same dataset partitioning strategy as [18], which randomly selects 250 makeup images and 100 non-makeup images as the test set, and the remaining images are used for training. We resize images to 256 × 256 in all experiments. We set $\lambda_{adv} = 1$, $\lambda_{cyc} = 10$, $\lambda_{per} = 0.005$, $\lambda_{local} = 1$, and $\lambda_{makeup} = 1$ for balancing the different loss functions. We set the numbers of RSTB, STL, and RAFB as 5, 6, and 6, respectively. We initialize ρ to 1 in RAFB. The optimizer of generators and discriminators is Adam [48] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The network is trained for 100 epochs with a batch size of 1 and a fixed learning rate of 0.0001. We implement HT-ASE with Pytorch and conduct all the experiments on an NVIDIA RTX 3090 GPU.

We compare the proposed HT-ASE with five state-of-the-art makeup transfer methods, including BeautyGAN [18], LADN [14], PSGAN [15], SCGAN [1] and SSAT [2]. We summarize the functions of the above makeup transfer methods in Table I where "Misalignment" indicates the robustness of makeup transfer in the case of large spatial misalignment between the source and referecne faces, "Detail" indicates the accurate transfer of the high-quality makeup style details, "Shade" and "Part" indicate the flexibility of shade-controllable transfer and partial transfer for lip, eye, and skin, respectively. As we can see from Table I that the proposed HT-ASE has the best ability and flexibility among all methods and is able to satisfy the needs in realistic application scenarios and handle the details of makeup styles.

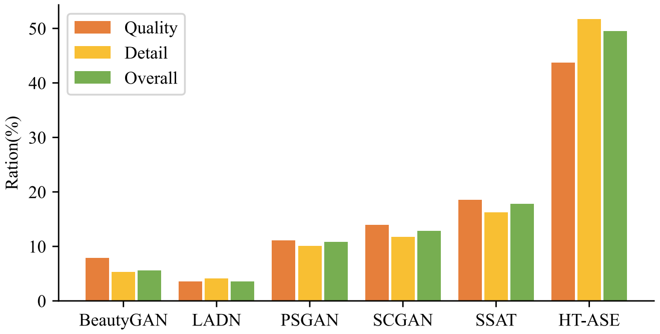


Fig. 7. User study results (ratio (%) selected as the best). "Quality", "Detail", and "Overall" denote the three aspects for evaluation: visual quality, the fidelity of makeup style details, and overall performance.

B. User Study

We conduct the user study between the proposed HT-ASE and five state-of-the-art makeup transfer methods. For the compared methods, we use their published pre-trained models to ensure fairness in quantitatively evaluating the visual quality and the precision of makeup details. We randomly select 10 non-makeup images and 15 references images from the MT test results, which contain both heavy and light makeup styles with good alignment, large misalignment, and facial occlusion. 30 volunteers (18 females and 12 males) aged from 18 years old to 35 years old participate in the user study and choose the best result from the images generated by the above-mentioned six methods according to visual quality, the preservation of makeup details with multiple colors and shapes, and overall performance. As shown in Fig. 7, our HT-ASE outperforms other compared methods, especially in the preservation of makeup details with multiple colors and shapes.

C. Quantitative Comparison

1) Expression/Pose Preservation: Makeup transfer requires that only the makeup styles can be transferred, and the expression/pose between the source face image and generated face image should be consistent. Therefore, we follow the evaluation strategy of PSGAN [15] to prove the preservation of expression and pose, which validates the landmark preservation of the source face image $\mathbf{A}$ and generated face image $\hat{\mathbf{A}}$ by cosine similarity metric. Specificaly, we use the Dlib 68 landmark detection algorithm to detect facial landmarks $\mathbf{l_A} = [x_{\mathbf{A}}^1, y_{\mathbf{A}}^1, x_{\mathbf{A}}^2, y_{\mathbf{A}}^2, \ldots, x_{\mathbf{A}}^t, y_{\mathbf{A}}^t]$ and $\mathbf{l}_{\hat{\mathbf{A}}} = [x_{\hat{\mathbf{A}}}^1, y_{\hat{\mathbf{A}}}^1, x_{\hat{\mathbf{A}}}^2, y_{\hat{\mathbf{A}}}^2, \ldots, x_{\hat{\mathbf{A}}}^t, y_{\hat{\mathbf{A}}}^t]$ of face images $\mathbf{A}$ and $\hat{\mathbf{A}}$, where $x_{\mathbf{A}}^d$ and $y_{\mathbf{A}}^d$ indicate the horizontal and vertical coordinates of

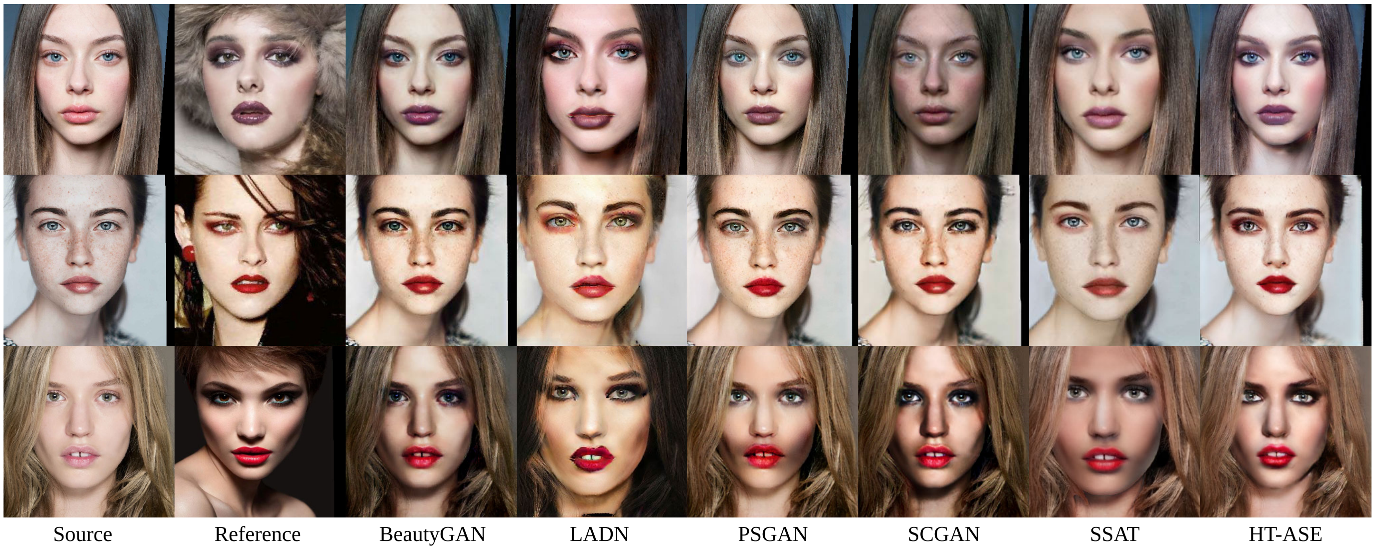


Fig. 8. Qualitative comparison with state-of-the-art methods over frontal faces and neutral expressions. Our proposed HT-ASE generates the most precise transferred result with the desired makeup style details.

TABLE II

QUANTITATIVE COMPARISON WITH EXISTING METHODS

Method	HT-ASE	BeautyGAN [18]	SCGAN [1]	PSGAN [15]	SSAT [2]	LADN [14]
cos_sim ↑	**0.9999**	0.9998	**0.9999**	0.9996	0.9990	0.9989
ArcFace ↑	**0.9416**	0.9396	0.9406	0.9383	0.9346	0.8442
FID ↓	**79.61**	83.50	93.22	79.96	89.94	96.03

TABLE III

QUANTITATIVE RESULTS OF THE ABILITY OF IDENTITY PRESERVING. OPTIMAL AND SUBOPTIMAL RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINE

Method	HT-ASE	BeautyGAN [18]	PSGAN [15]	SCGAN [1]	LADN [14]	SSAT [2]
FID ↓	$\underline{63.71}$	73.79	67.27	**60.91**	88.50	64.40
cos_sim ↑	**0.9998**	0.9996	$\underline{0.9997}$	0.9996	0.9990	0.9988

the d-th landmark in the source face image $\mathbf{A}$, $x^d_{\hat{\mathbf{A}}}$ and $y^d_{\hat{\mathbf{A}}}$ indicate the horizontal and vertical coordinates of the d-th landmark in the generated face image $\hat{\mathbf{A}}$. Then, we adopt the cosine similarity cos_sim to calculate the positional similarity of $\mathbf{l_A}$ and $\mathbf{l}_{\hat{\mathbf{A}}}$, which is defined as

$$\mathrm{H}_{cos\_sim} = \frac{\mathbf{l_A}\mathbf{l}_{\hat{\mathbf{A}}}{}^T}{\|\mathbf{l_A}\|\,\|\mathbf{l}_{\hat{\mathbf{A}}}\|} \tag{12}$$

The values of H_{cos_sim} for the proposed method and baselines are shown in Table II, which shows that our method can preserve the expression and pose of the source face image to the maximum extent.

2) Identity Preservation: To verify the ability of the proposed method for the preservation of face identities, we follow the evaluation strategy of PSGAN [15], which uses ArcFace [49] metric to compute the similarity between the source face image and the makeup transfer result. As shown in Table II, the proposed method achieves an average similarity of 0.9416 on the MT dataset [18], which is the best result among all methods. These results indicate that the proposed HT-ASE can preserve face identity well.

3) Quality and Authenticity: We calculate the FID [50] score between the makeup transfer results and reference images to evaluate the quality and authenticity of our method. As shown in Table II, the proposed method achieves the lowest FID score among all methods, which shows the capability of ensuring the quality and authenticity of the generated images.

4) Identity Preserving in Makeup Removal: To verify the ability of makeup removal, we compare our method with existing approaches in terms of identity preservation with the FID metric and cosine similarity metric in Table III. We achieves the optimal cos_sim score, which shows that our method can preserve the expression and pose of the source face image to the maximum extent. However, we acknowledge that the FID metric only focuses on the global style and is not sensitive to local details. To complement the quantitative evaluation, we provide qualitative comparisons in Fig. 10. While SCGAN achieves the optimal FID score, we observe a noticeable color difference in the eyes and face regions of the generated results, leading to unsatisfactory subjective effects. The reason is that SCGAN represents the three areas of makeup style details by three separate one-dimensional vectors, resulting in they are not unified. In contrast, our method generates more natural and visually appealing makeup removal results across various poses and makeup style details, including eye shadows, foundation, and lip gloss. By combining both quantitative and qualitative evaluations, we believe the performance of our method performance is well-demonstrated, providing a more comprehensive understanding of its effectiveness in preserving identity while achieving realistic and visually pleasing results.

D. Qualitative Comparison

1) Neutral Expressions: Fig. 8 shows the qualitative comparison between the proposed HT-ASE and other state-of-the-art methods on frontal images with neutral expressions and no obvious spatial misalignment. The results of

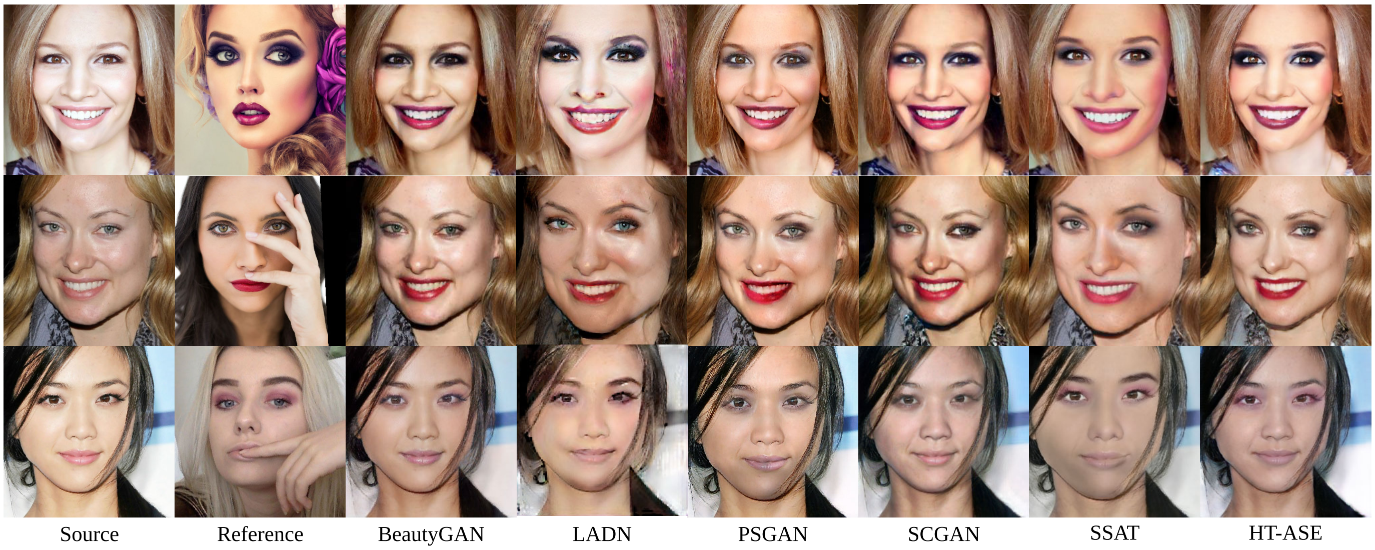


Fig. 9. Qualitative comparison with state-of-the-art methods under large spatial misalignment and partial facial occlusion. The proposed HT-ASE generates the most exquisite details (e.g., the shape and color of eye shadows), the most natural color, and the fewest artifacts.

BeautyGAN [18] and PSGAN [15] only have general makeup styles and lose many makeup details, such as the colors and shapes of eye shadows. LADN [14] is able to transfer makeup details. However, the transferred results are visually unacceptable since the colors of hair and makeup styles are inaccurate. Since SCGAN [1] represents makeup styles with spatially invariant 1D-vectors, it cannot distinguish different makeup regions. The first example shows a reference makeup style with a large color discrepancy between two facial regions (skin and eyes), which makes the makeup colors of two regions completely confused in the transferred result of SCGAN. Although the transferred results generated by SSAT [2] are natural, the colors of the makeup style details generated by SSAT tend to be lighter than the reference face images. Besides, it also loses much person-specific face identity information, such as the reflection and texture of source lips, which results in visual ambiguity. Moreover, since existing methods neglect to preserve the semantic information of face identity and fail to establish an accurate semantic correspondence between two faces, these methods suffer from the color bleeding problem at the lip edges. On the contrary, our proposed HT-ASE designs a makeup context extractor to learn context-aware makeup features that encode the complicated colors and shapes of the makeup styles and a face identity extractor to learn identity features with rich visual textures and semantic information, which guarantee the generation of the most realistic images with highly similar source face identity and reference makeup details.

2) Spatial Misalignment and Facial Occlusion: We compare the proposed HT-ASE with the state-of-the-art makeup transfer methods to test the robustness against large spatial misalignment and facial occlusion. The first row of Fig. 9 shows the makeup transfer results of the two face images with large spatial misalignment, where the compared methods can not accurately transfer makeup style details, including foundation color, the color and region of lipstick, shapes and colors of eye shadows and blush on cheeks, highlights on nose and skin. In contrast, our HT-ASE synthesizes vivid images with the same makeup styles as the reference face image, including natural illuminations and rich details mentioned above. The second and third rows show the makeup transfer results for the two face images with partial facial occlusions, such as hair occlusions in the original images, and hand occlusions in the reference images. BeautyGAN generates a general makeup style without partial eye shadows. LADN, PSGAN, SCGAN, and SSAT mistakenly transfer the features of hand and shadows in makeup regions, resulting in significant artifacts in transferred results, which demonstrate that these methods usually fail to repair the occluded regions. In contrast, our proposed method has demonstrated the capability to adaptively achieve semantic alignment, even under large occlusion cases. As a result, our proposed HT-ASE generates accurate and visually pleasing makeup transfer results, effectively handling the occlusion challenges presented in the experiment. On the one hand, existing methods lack the ability to learn context-aware makeup features that encode the complicated colors and shapes of the makeup styles and identity features with rich visual textures and semantic information. On the other hand, existing methods fail to establish an accurate semantic correspondence between two faces, resulting in makeup details being transferred to the wrong semantic positions and the inability to repair the occluded facial region. Therefore, we design a third sub-network called a spatially similarity-aware fusion network to handle large spatial misalignment and facial occlusion, which introduces a spatially-adaptive layer-instance normalization with attention-guided spatial embeddings to perform semantic alignment between the makeup and identity features under the correlation constraints of the identity features and then adaptively fuse the aligned makeup features and identity features to generate the transferred results.

3) Makeup Removal: In addition to makeup transfer, the proposed HT-ASE satisfies the requirements of an excellent makeup removal method, which removes the makeup style

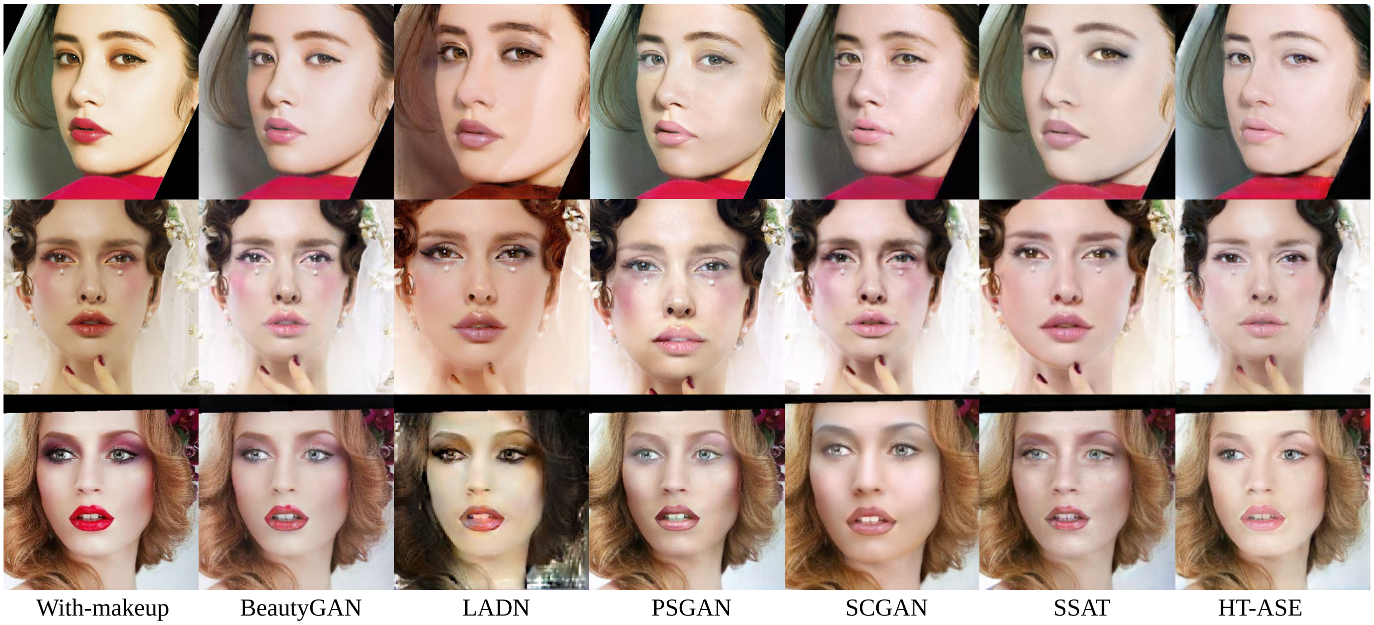


Fig. 10. Visualization results of makeup removal. Our HT-ASE generates natural and clean makeup removal results while preserving the face identities.

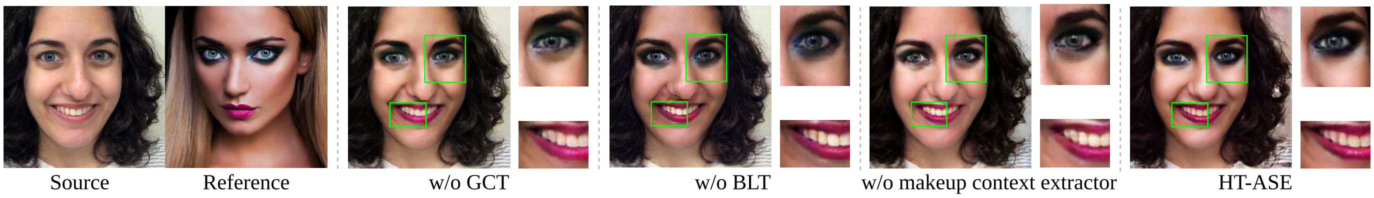


Fig. 11. Ablation studies of components in a makeup context extractor. "w/o GCT" denotes the makeup context extractor without a global context transformer (GCT). "w/o BLT" denotes makeup context extractor without a block-wise local transformer (BLT). "w/o makeup context extractor" denotes using common makeup extractor to replace our proposed makeup context extractor, such as those used in SCGAN [1] and SSAT [2]. We evaluate their effectiveness in the case of transferring makeup styles with misaligned expressions. Green boxes mark the comparisons of makeup style details.

while keeping the face identity intact after makeup removal. Fig. 10 shows the comparison of our HT-ASE with BeautyGAN [18], LADN [14], PSGAN [15], SCGAN [1], and SSAT [2], which demonstrates that our proposed HT-ASE method generates the most natural and precise makeup removal results even for various poses and makeup styles details, including the eye shadows, foundation, blush on cheeks and lip gloss. However, LADN and SCGAN cause severe artifacts in generated images, which are visually unrealistic. BeautyGAN, PSGAN, and SSAT do not remove the lip gloss and blush well.

E. Ablation Studies

1) Makeup Context Extractor: The makeup context extractor with a global context transformer (GCT) and a block-wise local transformer (BLT) is designed to aggregate the high-level context and low-level detail features of the makeup styles, which obtains the context-aware makeup features that encode the complicated colors and shapes of the makeup styles. As shown in Fig. 11, we verify the effectiveness of the makeup context extractor by conducting three ablation studies. The images in the first column are the source face image and the reference face image, which show a case of transferring makeup styles with misaligned expressions and poses. GCT uses non-local interactions to integrate high-level context features into low-level detail features. Without it, the shape of eye shadows in the transferred result (the third column) is incomplete, and the color of skin and lips are lighter than the reference. BLT uses low-level detail features to render high-level context features through local interactions. Without it, the color of eye shadows in the transferred result (the second column) is lighter than the reference. When using common makeup extractor to replace our proposed makeup context extractor, such as those used in SCGAN [1] and SSAT [2] that only contain residual convolution blocks, the generated results are unsatisfactory, e.g., the shape of eye shadows is incomplete and the highlights on skins are smooth or lost. The proposed makeup context extractor overcomes the above problems and obtains a precise makeup transfer result, which is highly consistent with the reference face image, especially with more makeup details. In addition, we conduct a qualitative comparison with various ablation studies on makeup context extractor. We calculate the FID score between the makeup transfer results and reference images to assess the quality and authenticity of our proposed various modules. As shown

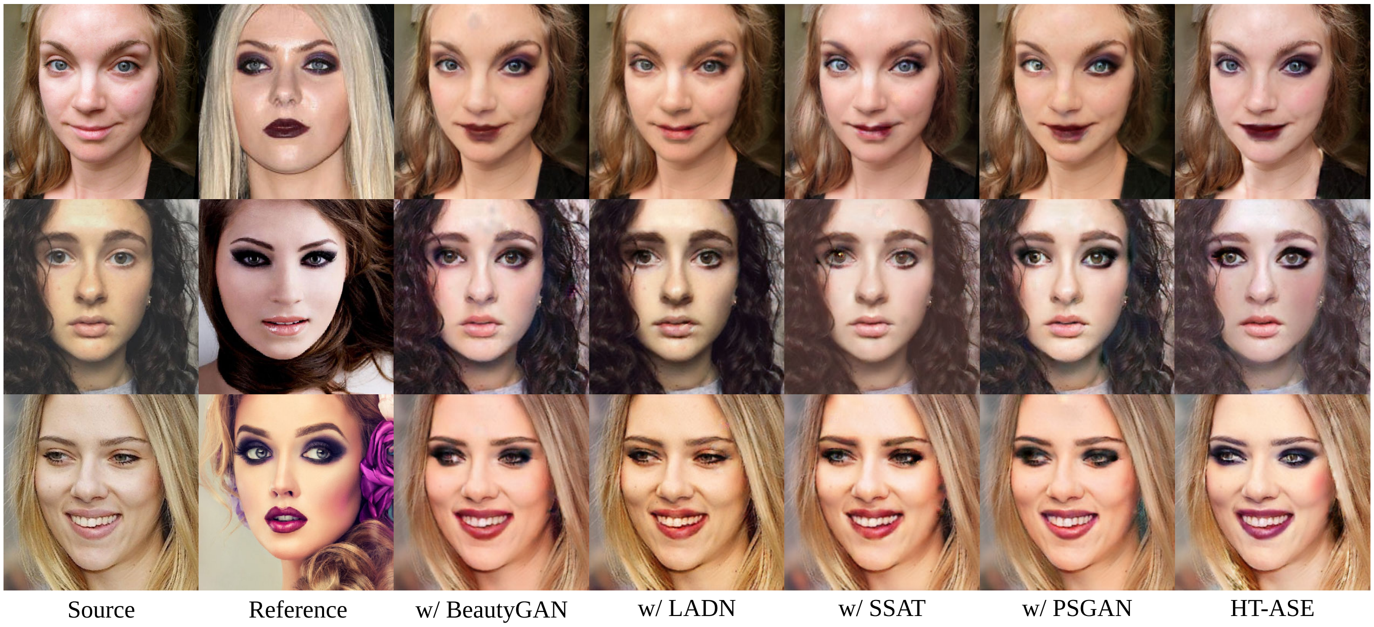


Fig. 12. Ablation studies on the face identity extractor. We evaluate the face identity extractors in existing methods for makeup style transfer.

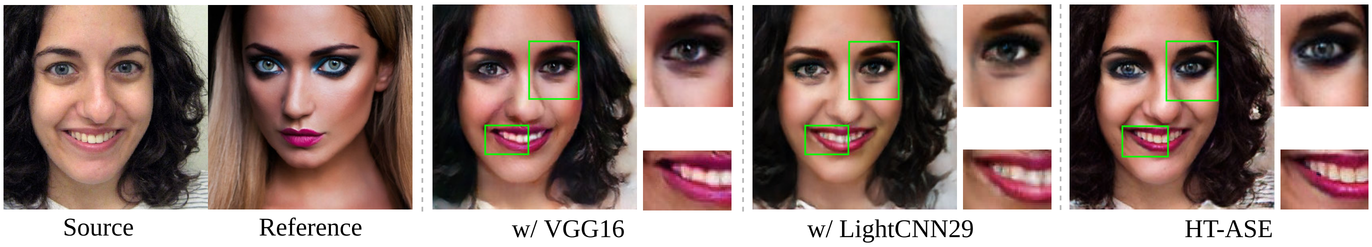


Fig. 13. Ablation studies on the structure of the face identity extractor. “w/ VGG16” and “w/ LightCNN29” denote the face identity extractors are VGG16 and LightCNN29, respectively. We evaluate these face identity extractors in the case of transferring makeup style with misaligned expressions. Green boxes highlight the makeup style details.

TABLE IV
ABLATION STUDIES ON THE COMPONENTS OF MAKEUP CONTEXT EXTRACTOR

Method	HT-ASE	w/o GCT	w/o BLT	w/o makeup context extractor
FID ↓	**79.61**	85.32	85.71	88.26

TABLE V
ABLATION STUDIES ON THE FACE IDENTITY EXTRACTOR

Method	HT-ASE	w/ BeautyGAN	w/ LADN	w/ SSAT	w/ PSGAN
FID ↓	**79.61**	83.44	89.90	90.36	84.30

in Table IV, the design of the complete makeup context extractor achieves the lowest FID score, which demonstrates the capability of the proposed makeup context extractor in ensuring the quality and authenticity of the generated images.

2) Face Identity Extractor: To demonstrate the superiority of our proposed face identity extractor over existing methods, we conduct extensive experimental evaluations. In Table V, our extractor achieves the best FID metric of 79.13, highlighting its ability to preserve face identity during makeup transfer. Additionally, a qualitative comparison in Fig. 12 demonstrates that existing extractors fail to preserve person-specific identity in crucial areas, while our method retains spatial information, resulting in more accurate makeup transfer with preserved makeup details. Our face identity extractor plays a crucial role in maintaining authenticity and realism, as confirmed by both quantitative and qualitative evaluations. Furthermore, we compare our face identity extractor with common VGG16 [47] and LightCNN29 [51] to verify the ability to extract identity information. In Fig. 13, we present qualitative results using different face identity extractors. The results show that neither VGG16 nor LightCNN29 can adequately preserve the visual texture information of facial regions such as teeth, lips, and eyes. Additionally, the transferred makeup styles lack detail, such as eyeshadow shapes and lipstick areas. This limitation arises because the architecture of VGG16 has a small receptive field and cannot capture fine-grained makeup details, while LightCNN29 struggles to effectively capture complex spatial semantic information of face identity. In contrast, our HT-ASE retains the complete structure of eye shadows and preserves many visual textures specific to the person's face identity, resulting in visually realistic transferred results. Furthermore, the proposed identity extractor network

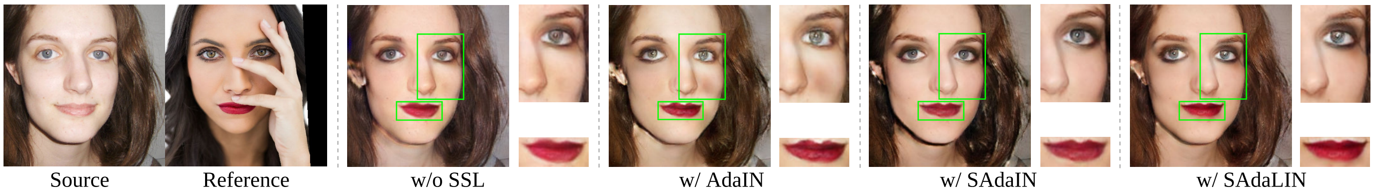


Fig. 14. Ablation studies on the components in spatially similarity-aware fusion network (SSFNet). "w/o SSL" indicates SSFNet does not have an SSL module to learn the semantic similarity matrix between two faces. "w/ AdaIN" indicates AdaIN [43] is performed in the feature fusion phase instead of SAdaLIN. "w/ SAdaIN" indicates that AdaIN with attention-guided spatial embeddings is performed in the feature fusion phase instead of SAdaLIN. We evaluate their effectiveness in the case of handling a reference face image with partial facial occlusion. Green boxes mark the comparisons of makeup style details.

TABLE VI
ABLATION STUDIES ON THE STRUCTURE OF THE FACE IDENTITY EXTRACTOR

Method	HT-ASE	w/ VGG16	w/ LightCNN29
FID ↓	**79.61**	86.30	87.37

TABLE VII
ABLATION STUDIES ON DIFFERENT COMPONENTS OF SPATIALLY SIMILARITY-AWARE FUSION NETWORK

Method	HT-ASE	w/o SSL	w/ AdaIN	w/ SAdaIN
FID ↓	**79.61**	91.33	86.31	88.38

TABLE VIII
MODEL COMPLEXITY ANALYSIS.THE FLOPS (G), MEMORY USAGE (M), AND INFERENCE TIME (SECOND) FOR PROCESSING IMAGES WITH A SIZE OF 256 × 256 ARE PRESENTED.OPTIMAL AND SUBOPTIMAL RESULTS ARE HIGHLIGHTED BY BOLD AND UNDERLINE

Method	FLOPs (G)↓	Memory Usage (M)↓	Inference Time (s)↓
SCGAN	1115.19	236	0.1011
PSGAN	**38.9**	**150**	**0.0185**
LADN	<u>728.28</u>	1032	<u>0.0436</u>
SSAT	737.10	268	0.0440
HT-ASE	754.68	<u>223</u>	0.1620

exhibits the best performance in evaluation scores by integrating multi-scale identity semantic features and aggregating identity-relevant features at different scales. Table VI reports the FID score comparison between the makeup transfer results and reference images, assessing the quality and authenticity of various modules. Our HT-ASE achieves the lowest FID score, showcasing its superiority in generating high-quality and realistic images. Both qualitative and quantitative results demonstrate the superiority of our HT-ASE, leading us not to adopt VGG16 or LightCNN29.

3) Spatially Similarity-Aware Fusion Network: The spatially similarity-aware fusion network (SSFNet) includes a semantic similarity learning (SSL) module and four spatially-adaptive layer-instance normalization (SAdaLIN) modules, which is designed to generate the transferred results even with large spatial misalignment and facial occlusion. As shown in Fig. 14, we verify the effectiveness of SSFNet by conducting three ablation studies. The images in the first column depict source and reference face images, showing a case of transferring makeup styles with facial occlusion. "w/o SSL" indicates SSFNet does not have an SSL module to learn the semantic similarity matrix between the identity features. "w/ AdaIN" refers to using AdaIN [43] in the feature fusion phase instead of SAdaLIN, while "w/ SAdaIN" indicates using AdaIN with attention-guided spatial embeddings in the feature fusion phase instead of SAdaLIN. All transferred results fail to repair occluded regions and produce artifacts, while our proposed SAdaLIN preserves makeup style details and repairs occluded regions. Furthermore, we conduct a qualitative comparison with various ablation studies on SAdaLIN and calculate the FID score between the makeup transfer results and reference images to assess the quality and authenticity of our proposed modules. Table VII demonstrates that the complete SAdaLIN design achieves the lowest FID score, indicating its capability to ensure high-quality and authentic generated images. Moreover, we conduct a qualitative comparison with various ablation studies on SSFNet. We calculate the FID score between the makeup transfer results and reference images to assess the quality and authenticity of our proposed various modules. The design of the complete SSFNet achieves the lowest FID score in Table VII, which demonstrates the capability of ensuring the quality and authenticity of the generated images.

4) Computing Cost Comparisons: Table VIII presents a comprehensive comparison of our model with existing methods in terms of computational costs, number of parameters, and inference time. The experimental results demonstrate that our method does not substantially increase model complexity compared to other methods, despite its excellent performance. This finding strongly supports the effectiveness of our proposed method, as shown in Figs. 8-10.

F. Characteristics of Makeup Transfer

1) Partial Makeup Transfer: Since the makeup features have a context-aware characteristic, HT-ASE achieves partial makeup transfer. Given a source face image $\mathbf{X}$ and a reference face image $\mathbf{Y}_p$ of which we want to transfer for part p, we can obtain the identity features $\mathbf{X}^c, \mathbf{Y}_p^c$ and the makeup feature $\mathbf{X}^s, \mathbf{Y}_p^s$ by feeding the images to the face identity extractor and makeup context extractor. Under the guidance of face parsing masks $\mathbf{M}_X \in [0,1]^{H\times W}, \mathbf{M}_{Y_p} \in [0,1]^{H\times W}$, which are obtained according to the face parsing method in [39] and expanded along the channel dimension, the partial transfer feature map $\hat{\mathbf{Y}}_p^c$ is calculated as:

$$\hat{\mathbf{Y}}_p^c = \mathbf{M}_{Y_p} \odot \mathbf{Y}_p^s + (1 - \mathbf{M}_{Y_p}) \odot \mathbf{X}^s \qquad (13)$$

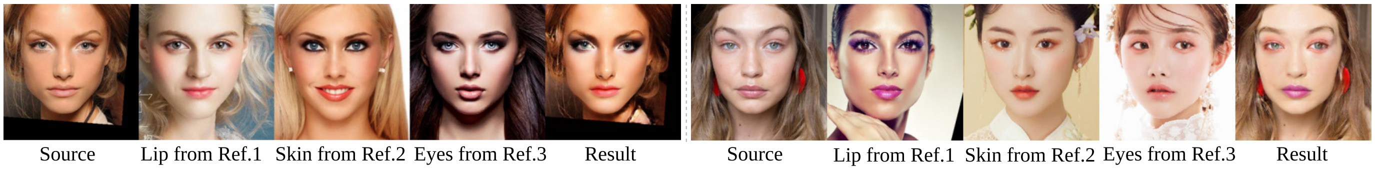


Fig. 15. Illustration of partial makeup transfer. The last column is the partial makeup transfer results, which receive personal identity from the first column, the lips style from the second column, the eyes style from the third column, and the face style from the fourth column.

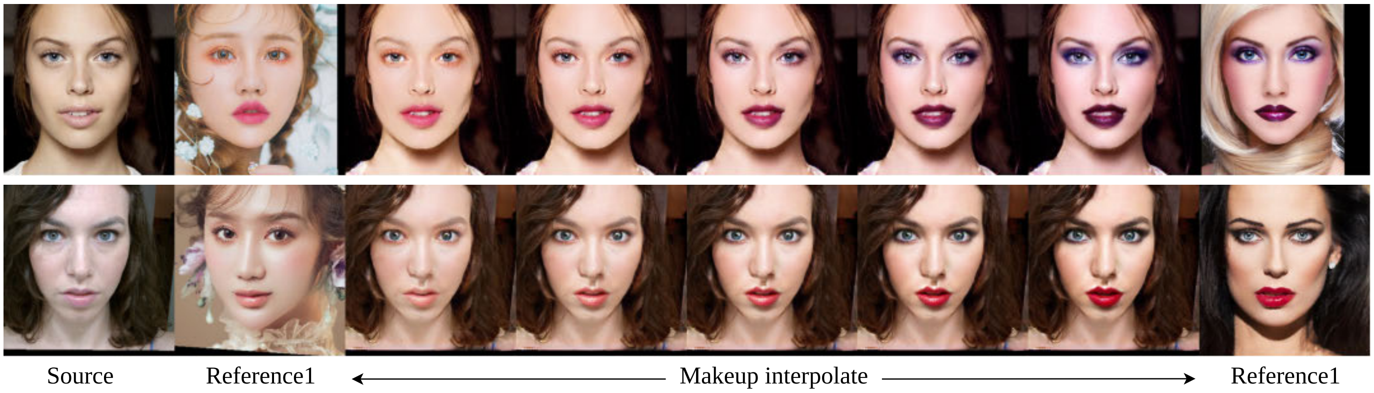


Fig. 16. Illustrations of intensity controllable makeup transfer. The two rows show the transferred results of makeup transfer effect from light to heavy.

where ⊙ denotes the Hadamard product. Fig. 15 shows two realistic and natural results of partial makeup transfer, where the makeup style of skin, eyes, and lips are from Ref1, Ref2, and Ref3, respectively.

2) Intensity Controllable Makeup Transfer: Since makeup features have a context-aware characteristic, HT-ASE achieves intensity controllable makeup transfer. We add a modulation factor $\alpha \in [0, 1]$ to makeup features, which makes it easy to interpolate makeup style from multiple references. In this experiment, we test to obtain a new makeup feature map by interpolating two makeup features from two references $\mathbf{Y}_{img1}$ and $\mathbf{Y}_{img2}$. The new makeup feature map $\hat{Y^s}$ is calculated by

$$\hat{\mathbf{Y}}^s = \alpha \mathbf{Y}^s_{img1} + (1-\alpha)\mathbf{Y}^s_{img2} \tag{14}$$

Fig. 16 shows a series of results from one source image and two reference images with different levels of contributions to makeup styles. The makeup style of generated images transits from Reference 1 to Reference 2 by adjusting the modulation factor α from 0 to 1 in increments of 0.2.

3) Limitations: Our HT-ASE can be improved in two aspects. First, facial images usually have high resolutions in practical application scenarios, such as 512 × 512 or even 1024 × 1024. Therefore, it is necessary to ensure that the makeup transfer algorithm is still fast and effective at high resolution. Second, our method can transfer the color and shape of makeup style but fails to handle the stickers on the faces, which are treated as identity features. As shown in Fig. 17, we try makeup transfer and removal with source face images (the first column) and reference face images (the second column), and the results indicate that the stickers on the reference face images are retained as information related to the face identity.

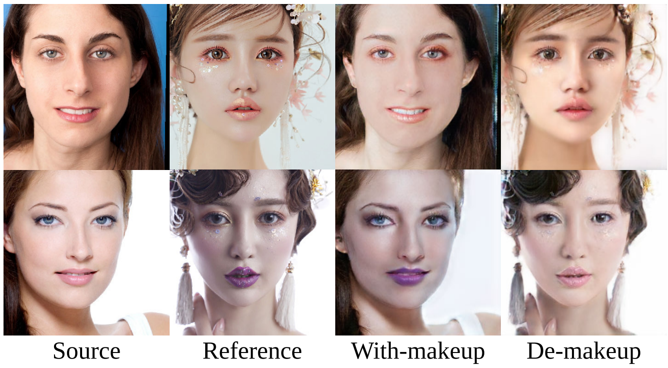


Fig. 17. Limitation of our method. Although we can successfully transfer the base makeup, special effects cannot be correctly assigned.

V. Conclusion and Future Work

This paper proposes Hybrid CNN-transformers with Attention-guided Spatial Embeddings (named HT-ASE) for makeup transfer and removal. Specifically, a makeup context extractor adopts makeup context global-local interactions to aggregate the high-level context and low-level detail features of the makeup styles, respectively, which are further integrated with multi-scale features to obtain the context-aware makeup features that encode the complicated colors and shapes of the makeup styles. A face identity extractor adopts the face identity local interaction with two block-wise local transformers to aggregate identity-relevant features at different scales into identity semantic features, which are further integrated with multi-scale features to refine the identity features. A spatially similarity-aware fusion network introduces a spatially-adaptive layer-instance normalization with attention-guided spatial embeddings to perform semantic alignment between the makeup and identity features under the correlation constraints of the identity features and then adaptively fuse the aligned makeup features and identity features to generate the transferred results even with large spatial misalignment and facial occlusion. Extensive experiments demonstrate that our proposed method outperforms the state-of-the-art methods. For future work, we consider exploring semantic segmentation methods to handle face stickers in makeup styles. Besides, we consider studying higher-resolution makeup transfer for a broader range of realistic application scenarios.

References

[1] H. Deng, C. Han, H. Cai, G. Han, and S. He, "Spatially-invariant style-codes controlled makeup transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6545–6553.

[2] Z. Sun, Y. Chen, and S. Xiong, "SSAT: A symmetric semantic-aware transformer network for makeup transfer and removal," in *Proc. AAAI*, 2022, vol. 36, no. 2, pp. 2325–2334.

[3] Y. Lyu, J. Dong, B. Peng, W. Wang, and T. Tan, "SOGAN: 3D-aware shadow and occlusion robust GAN for makeup transfer," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3601–3609.

[4] T. Nguyen, A. T. Tran, and M. Hoai, "Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13300–13309.

[5] Z. Wan, H. Chen, J. An, W. Jiang, C. Yao, and J. Luo, "Facial attribute transformers for precise and robust makeup transfer," in *Proc. WACV*, 2022, pp. 1717–1726.

[6] S. Liu et al., "PSGAN++: Robust detail-preserving makeup transfer and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8538–8551, Nov. 2022.

[7] W.-S. Tong, C.-K. Tang, M. S. Brown, and Y.-Q. Xu, "Example-based cosmetic transfer," in *Proc. 15th Pacific Conf. Comput. Graph. Appl. (PG)*, Oct. 2007, pp. 211–218.

[8] D. Guo and T. Sim, "Digital face makeup by example," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 73–79.

[9] C. Li, K. Zhou, and S. Lin, "Simulating makeup through physics-based manipulation of intrinsic image layers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4621–4629.

[10] *TikTok*. [Online]. Available: https://www.tiktok.com/en/

[11] *MeiTu*. [Online]. Available: https://www.meitu.com/en

[12] W. Wang, Y. Fu, X. Qian, Y.-G. Jiang, Q. Tian, and X. Xue, "FM^{2}U-Net: Face morphological multi-branch network for makeup-invariant face verification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5729–5739.

[13] S. Hu et al., "Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14994–15003.

[14] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang, "LADN: Local adversarial disentangling network for facial makeup and de-makeup," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10480–10489.

[15] W. Jiang et al., "PSGAN: Pose and expression robust spatial-aware GAN for customizable makeup transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5193–5201.

[16] I. Goodfellow et al., "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 1–14.

[17] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. ECCV*, 2018, pp. 35–51.

[18] T. Li et al., "BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 645–653.

[19] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "BeautyGlow: On-demand makeup transfer framework with reversible generative network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10034–10042.

[20] Y. Li, H. Huang, J. Cao, R. He, and T. Tan, "Disentangled representation learning of makeup portraits in the wild," *Int. J. Comput. Vis.*, vol. 128, nos. 8–9, pp. 2166–2184, Sep. 2020.

[21] J. Xiang, J. Chen, W. Liu, X. Hou, and L. Shen, "RamGAN: Region attentive morphing GAN for region-level makeup transfer," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022, pp. 719–735.

[22] Y. Gao et al., "Wallpaper texture generation and style transfer based on multi-label semantics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1552–1563, Mar. 2022.

[23] L. Fu, H. Yu, F. Juefei-Xu, J. Li, Q. Guo, and S. Wang, "Let there be light: Improved traffic surveillance via detail preserving night-to-day transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8217–8226, Dec. 2022.

[24] H. Yan, H. Zhang, J. Shi, and J. Ma, "Texture brush for fashion inspiration transfer: A generative adversarial network with heatmap-guided semantic disentanglement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2381–2395, May 2023.

[25] Z. Chen, W. Wang, E. Xie, T. Lu, and P. Luo, "Towards ultra-resolution neural style transfer via thumbnail instance normalization," in *Proc. AAAI*, 2022, vol. 36, no. 1, pp. 393–400.

[26] H. Chen et al., "Quality evaluation of arbitrary style transfer: Subjective study and objective metric," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3055–3070, Jul. 2023.

[27] P. Zhou, L. Xie, B. Ni, L. Liu, and Q. Tian, "HRInversion: High-resolution GAN inversion for cross-domain image synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2147–2161, May 2023.

[28] J. Yang, F. Guo, S. Chen, J. Li, and J. Yang, "Industrial style transfer with large-scale geometric warping and content preservation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7824–7833.

[29] Z. Li et al., "SDTP: Semantic-aware decoupled transformer pyramid for dense image prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6160–6173, Sep. 2022.

[30] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12889–12899.

[31] Y. Deng et al., "StyTr^2: Image style transfer with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11316–11326.

[32] X. Wu, Z. Hu, L. Sheng, and D. Xu, "StyleFormer: Real-time arbitrary style transfer via parametric style composition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14598–14607.

[33] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[34] Z. Sun, F. Liu, W. Liu, S. Xiong, and W. Liu, "Local facial makeup transfer via disentangled representation," in *Proc. ACCV*, 2020, pp. 1–15.

[35] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[36] Y. Deng, F. Tang, W. Dong, W. Sun, F. Huang, and C. Xu, "Arbitrary style transfer via multi-adaptation network," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2719–2727.

[37] L. Chen et al., "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6298–6306.

[38] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–16, Aug. 2018.

[39] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. ECCV*, 2018, pp. 325–341.

[40] H. Nam and H.-E. Kim, "Batch-instance normalization for adaptively style-invariant neural networks," *Proc. NIPS*, vol. 31, 2018, pp. 1–10.

[41] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," 2019, *arXiv:1907.10830*.

[42] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[43] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[44] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[45] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[46] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.

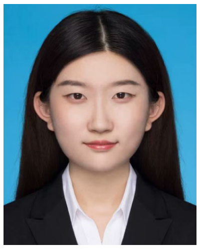
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[49] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[50] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. NIPS*, vol. 30, 2017, pp. 1–12.

[51] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.



Mingxiu Li received the B.S. degree from Inner Mongolia University, China, in 2020. She is currently pursuing the Ph.D. degree in computer science and technology with the Harbin Institute of Technology (HIT), Weihai. Her research interests include low-level vision, makeup transfer, and style transfer.

Wei Yu received the B.S. and M.S. degrees from the China University of Petroleum (East China), China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree in computer science and technology with the Harbin Institute of Technology (HIT), Weihai. His research interests include low-level vision, image super-resolution, and restoration.

Qinglin Liu received the B.S. degree from Yanshan University, China, in 2014, and the M.S. degree from Xidian University, China, in 2018. He is currently pursuing the Ph.D. degree in computer science and technology with the Harbin Institute of Technology (HIT), Weihai. His research interests include low-level vision, image segmentation, and image matting.

Zonglin Li received the B.S. degree from the Harbin Institute of Technology, Weihai, in 2017, and the M.S. degree from the University of Pittsburgh, USA, in 2019. He is currently pursuing the Ph.D. degree with the Faculty of Computing, Harbin Institute of Technology (HIT). His research interests include computer vision, computer graphics, and 3D reconstruction.

Ru Li (Member, IEEE) received the B.E. degree in electronic information engineering from the China University of Petroleum, Qingdao, China, in 2016, and the Ph.D. degree from the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2022. She was a Visiting Student Researcher with the University of Oxford. She is currently a Lecturer with the School of Computer Science and Technology, Harbin Institute of Technology (HIT), Weihai. Her research interests include image processing and computer vision.

Bineng Zhong received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2004, 2006, and 2010, respectively. From 2007 to 2008, he was a Research Fellow with the Institute of Automation and the Institute of Computing Technology, Chinese Academy of Sciences. From September 2017 to September 2018, he was a Visiting Scholar with Northeastern University, Boston, MA, USA. He is currently a Professor with the School of Computer Science and Engineering, Guangxi Normal University, China. His research interests include pattern recognition, machine learning, and computer vision.

Shengping Zhang (Member, IEEE) received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2013. He was a Post-Doctoral Research Associate with Brown University and Hong Kong Baptist University. He was a Visiting Student Researcher with the University of California at Berkeley. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai. He has authored or coauthored more than 50 research publications in refereed journals and conferences. His research interests include deep learning and its applications in computer vision.