Canonical Makeup Transfer

Xinyu Lin[1*]
linxinyu0327@gmail.com
Kun Zhou[2,3*]
zhoukun303808@gmail.com
Xiaoguang Han[1†]
hanxiaoguang@cuhk.edu.cn
Jiangbo Lu[2†]
jiangbo.lu@gmail.com

[1] SSE,
CUHK-Shenzhen,
Shenzhen, China.

[2] SmartMore Corporation,
Shenzhen, China.

[3] School of Architecture & Urban Planning,
Shenzhen University,
Shenzhen, China.

Abstract

Makeup transfer aims at adapting makeup styles from reference images to source non-makeup images, typically portraits. Despite significant progress in this area, challenges persist, particularly in effectively transferring makeup across faces with notable pose or expression variations. In this study, we introduce a canonical makeup transfer (CMT) approach to reduce the disparity between the source and reference style faces, while ensuring semantic consistency in the transformation process. By minimizing this transfer gap, our method facilitates robust and efficient interaction between source and style facial images. Additionally, we establish a comprehensive benchmark for evaluating the performance of state-of-the-art methods across diverse scenarios. Extensive experiments demonstrate that our approach achieves superior results in terms of transfer quality and artifact reduction.

1 Introduction

Makeup transfer, aimed at replicating makeup styles from a reference face to a source face, is essential for applications like virtual try-ons [3, 5, 9, 10, 19, 31]. This task transfers makeup styles, like color distribution, while simultaneously preserving background and identity traits such as geometry, expressions.

A key challenge lies in modeling semantic correspondences between source and reference images, especially under facial spatial misalignment [4, 11, 28, 29]. Early methods like PSGAN [11] and FAT [23] use cross-attention for pixel-level correspondence but do not explicitly align features/images of source-reference pairs, relying only on attention to transfer between misaligned faces. More recently, EleGANt [29] improves on this by explicitly aligning features before attention, to minimize the spatial distance for semantic correspondences. In detail, EleGANt performs style transfer in an instance-dependent way by directly aligning the style image to the source image (we call it 'one-side alignment'), as illustrated in

*Equal contribution
†Corresponding author

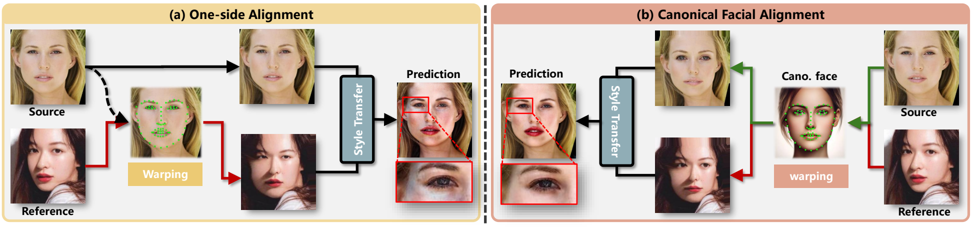

Figure 1: Comparison between conventional one-side alignment and our novel canonical facial alignment schemes. While the former often leads to noisy makeup outputs (severe color distortion below the left eye) caused by significant misalignment and semantic inconsistencies, our proposed CMT establishes semantic-invariant space and substantially diminishes such alignment disparities, resulting in more precise transferred faces with fewer artifacts.

Fig. 1, and then applies style transformation on the aligned pair. However, there are several ***potential issues***: (1) Style information loss produced by spatial warping, especially for the large pose and facial expression variances (spatial misalignment). (2) In such an instance-dependent warping scheme, it is difficult to robustly predict the semantic-varying output (as shown in Fig. 4, the semantic meaning of a fixed pixel location varies).

To address these challenges, we propose a Canonical Makeup Transfer (CMT) strategy. As shown in Fig.1, CMT defines a canonical face that minimizes misalignment across all training samples, and then aligns both source and style images to this optimized canonical face. CMT has two key advantages: (1) It creates a semantic-invariant space to enable semantically consistent source-style interaction, and reduce the information loss caused by inaccurate alignment; (2) The semantic priors and constraints yielded by canonical face reduce the prediction variances, leading to stable optimization. Overall, similar to normalization methods[8, 17], CMT normalizes pose and expression variations, achieving improved transfer results (shown in Fig. 4).

On the other hand, existing methods [13, 26, 28, 29] often form test sets by randomly selected test images from public dataset, lacking evaluation consistency and misalignment diversity, making thorough evaluation difficult. To address this issue, we design a detailed makeup benchmark by carefully categorizing test image pairs based on spatial distances, enabling comprehensive and fair model comparisons. Moreover, we argue that merely providing visual samples is insufficient for a comprehensive evaluation, as objective assessments are necessary. RamGAN [26] utilized SSIM for such assessments by comparing structural similarities between source and transferred results. However, changes in makeup style can alter local structures, reducing the accuracy of direct SSIM calculations between source and transferred images.

In conclusion, our contributions are summarized as follows:

- **Canonical makeup transfer.** We present a canonical makeup transfer approach to shorten the spatial misalignment between source and reference face pairs, resulting in effective semantic-consistent interaction and better information preservation.

- **Benchmarking and quantitative metrics.** We carefully design a fine-grained evaluation benchmark MT-Eval for fair and comprehensive evaluations under different scenarios. Moreover, we present the cyclic reconstruction metrics (in Sec. 4) for objective assessments.

- **SOTA performance.** Extensive experiments and user study demonstrate that our method achieves state-of-the-art performance in transferring makeup quality.

2 Related Work

Numerous methods [4, 6, 11, 13, 20, 24, 26, 29] have been proposed for makeup transfer. Early approaches [7, 12, 14, 18, 22, 27] relied on image processing techniques like facial feature detection, color transfer, and blending. These were later outpaced by deep learning methods, starting with Liu et al. [15], who introduced the first deep learning-based framework for part-to-cosmetic mapping and transfer.

Given the lack of paired data, many recent works adopt unsupervised GAN-based approaches. BeautyGAN [13] introduces global and local losses; PairedCycleGAN [1] builds on CycleGAN for example-based transfer; BeautyGlow [2] separates latent codes into makeup and non-makeup factors; and LADN [6] enables local detail transfer via multiple discriminators. SCGAN [4] encodes component-wise styles into spatial-invariant codes.

To address spatial misalignment, attention-based methods were introduced. PSGAN [11] pioneers Attentive Makeup Morphing for pixel-wise alignment. FAT [23] uses transformers to model semantic interactions. EleGANt [29] enables localized editing via shifted window attention. SSAT [20] proposes a symmetric transformer for joint transfer and removal, while RamGAN [26] introduces Region Attentive Matrices for fine-grained region-aware transfer.

However, these attention-based methods attempt to perform style transfer in an instance-dependent space: directly aligning the style image into the source image and performing cross-attention between the warped style and unwarped source image pair. In addition to information loss due to the large pose and facial expression variances, it is difficult to predict the semantic-varying outputs that align with source images. In contrast, we propose a canonical makeup transfer strategy for shortening the averaging distances between style and source images. Moreover, it performs a semantic-consistent interaction in a canonical facial space, significantly enhancing the robustness of our results.

3 Methodology

3.1 Problem Definition

Makeup transfer aims to transfer the makeup style from a reference image to a non-makeup source image, while preserving the identity, facial geometry, illumination, and other non-target regions. Here we set $\mathcal{X} \subset \mathbb{R}^{H\times W\times 3}$ and $\mathcal{Y} \subset \mathbb{R}^{H\times W\times 3}$ as the source image domain and the reference image domain, respectively. Please note that a matching pair of makeup and non-makeup images is not available; that is, the identities of the source and reference images are totally different.

Given a non-makeup sample $x \in \mathcal{X}$ and a makeup sample $y \in \mathcal{Y}$, the goal of the makeup transfer is to learn a mapping function $\mathcal{G} : x \to \tilde{y_x}$, where $\tilde{y_x} \in \mathcal{Y}$ possesses the makeup style of y and the identity of x.

3.2 Overview

Fig. 2 illustrates the framework of our approach, comprising a Facial Feature Encoder (FFE), Canonical Makeup Transfer (CMT), and a Facial Makeup Decoder (FMD): (a) The *Facial*

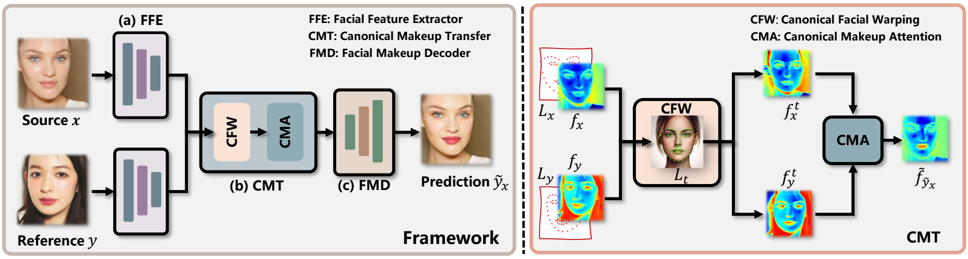

Figure 2: An outline of our newly introduced canonical makeup transfer framework comprising three key components: (a) A facial feature extractor (FFE) responsible for acquiring pyramid feature representations from both source and reference images. (b) The canonical makeup transfer (CMT) module, which initially aligns the features from both images into a canonical facial space (see Eq. 3) and subsequently employs a canonical makeup attention mechanism (CMA) to compute aggregated features. (c) The facial makeup decoder (FMD) is tasked with processing the aggregated features to generate the final outcome. (For description clarity, here we temporally drop the pyramid superscript $\{1,2,\cdots,M\}$ for features.)

Feature Extractor (FFE) encodes facial attributes and makeup styles into multi-scale pyramid feature maps from source and style inputs (x,y) (RGB images). (b) The *Canonical Makeup Transfer* firstly aligns feature pairs into a canonical facial space for spatial consistency, then fuses target styles (e.g., lips, eyes, skin) and source content using a makeup attention module. (c) The *Facial Makeup Decoder* progressively restores the final prediction from the fused pyramid features. Detailed designs are described in the following section.

3.3 Facial Feature Extractor (FFE)

Following previous works [4, 11, 29], we adopt a convolutional module to extract multi-scale facial information from the input source and style images. As shown in Fig. 3 (a), it consists of several residual blocks, where each residual block consists of several convolutional layers with a skip connection. Thus, our FFE module obtains deep (M levels) pyramid image features from both input source and style images (x,y):

$$f_{\{x\}}^{\{1,2,\cdots,M\}} = \mathrm{FFE}(x), \quad f_{\{y\}}^{\{1,2,\cdots,M\}} = \mathrm{FFE}(y), \tag{1}$$

3.4 Canonical Makeup Transfer (CMT)

Previous studies [29] typically conduct one-sided alignment between source and reference images, improving performance over earlier methods [4, 11, 23]. However, there are several drawbacks: (1) Warping strategy with large pose and expression differences may cause information loss. (2) Facial style transfer in a space with semantic variations further hinders the robust generation.

To address these issues, we propose the novel Canonical Makeup Transfer (CMT), which uses canonical facial warping to enable style-content attention in a semantic-invariant space for robust interaction, reducing spatial misalignment and enhancing information preservation. CMT includes: (1) Aligning source and style faces in our canonical facial space.

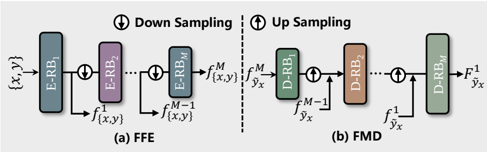

Figure 3: Illustration of our FFE and FMD modules. (a) FFE consists of M residual blocks (denoted as E-RB$_{\{1,2,\cdots,M\}}$), and the m-th block produces $f^m_{\{x,y\}}$. (b) FMD uses M similar residual blocks (D-RB$_{\{1,2,\cdots,M\}}$) to progressively fuse the transferred facial features $f^{\{1,2,\cdots,M\}}_{\tilde{y}_x}$.

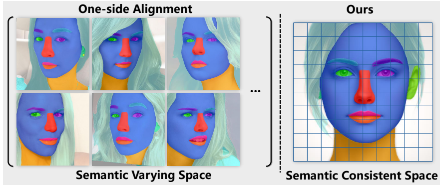

Figure 4: Illustrations of semantic space for one-side alignment vs. our canonical transfer strategy: one-side alignment operates in a semantically varying space, while our canonical facial space ensures semantic consistency.

(2) Employing cross-attention to transfer makeup information to source features within the canonical space.

Canonical Facial Warping (CFW). In order to reduce misalignment and improve the efficacy of style transfer, we introduce a canonical facial warping scheme. Illustrated in Fig. 2, our approach begins by optimizing a canonical face $L_t = \{l^1_t, l^2_t, \cdots, l^N_t\}$, where l^j_t denotes the j-th landmark of L_t, defined by N facial landmarks as follows:

$$\begin{aligned} L_t &= \min_L \sum_i \Phi(L, L_i), \\ \Phi(L, L_i) &= \frac{1}{N} \sum_{j=1}^{N} \mathrm{Dist}(l^j, l^j_i), \end{aligned} \tag{2}$$

where L_i denotes the i-th sample in training set. In this context, L_t represents the optimal face with minimal misalignment across the training samples (both source and style images are used here). Φ computes the average distance between each pair of face landmarks, and our canonical facial space is delineated by the landmarks within L_t.

Subsequently, we proceed to warp the multi-scale feature pairs $f^{\{1,2,\cdots,M\}}_{\{x,y\}}$ produced by the facial feature extractor into our canonical facial space. To do this, we employ the N landmarks as control points to deform $f^{\{1,2,\cdots,M\}}_{\{x,y\}}$ as follows:

$$f^{\{1,2,\cdots,M\}}_{\{x,y\},t} = TPS(f^{\{1,2,\cdots,M\}}_{\{x,y\}}, L_{\{x,y\}}, L_t), \tag{3}$$

where $f^{\{1,2,\cdots,M\}}_{\{x,y\},t}$ denotes the deformed source and style feature pairs in the defined canonical facial space, and TPS denotes the Thin Plate Splines [25], and L_t represents the optimized canonical face. As a result, the facial features $f^{\{1,2,\cdots,M\}}_{\{x,y\},t}$ are aligned with the canonical face according to the sparse guidance of landmark deformation ($L_{\{x,y\}} \to L_t$).

Canonical Makeup Attention (CMA). In CFW, we have warped multi-scale source and style image feature pairs into our canonical facial space. Here, we further describe our canonical makeup attention scheme to transfer makeup information from these warped feature pairs. And for description clarity, we temporally drop the pyramid superscript $\{1,2,\cdots,M\}$.

CMA utilizes a window-based QKV-cross-attention to aggregate dual-domain information from the warped source and style pyramid feature pairs. To begin with, a feature pair

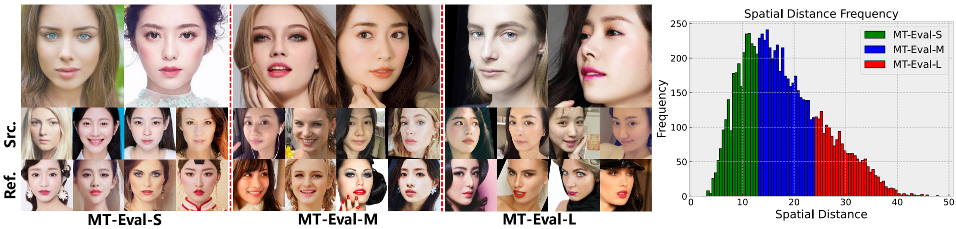

Figure 5: Illustrations of some sample images in MT-Eval. We divide the sub-datasets based on the spatial distance in Eq. 9. Note that green, blue, and red regions indicate samples in MT-Eval-S, MT-Eval-M and MT-Eval-L, respectively. "Src." and "Ref." refer to the source and reference images.

$f^t_{\{x,y\}}$ ('t' represents in canonical space) are reshaped into patches with a shape of $S^2 \times C$. For the patch pairs $s^t_{\{x,y\}}$, the cross-attentive matrix A and aggregation are achieved by:

$$\begin{aligned} &Q = s^t_x P^Q, \quad K = s^t_y P^K, \quad V = s^t_y P^V, \\ &A = \mathrm{SoftMax}\left(\frac{QK^T}{\sqrt{C}} + B\right), \\ &Y = AV, \end{aligned} \tag{4}$$

where $P^Q, P^K, P^V \in R^{\{C\times C\}}$ denote the learnable feature projection parameters, $Q, K, V \in R^{\{S^2\times C\}}$ represent the projected features of patch pairs $s^t_{\{x,y\}}$ and B represents a relative position encoding. Next, all the aggregated patches are summed and reshaped into 2D feature $f^{att}_{\tilde{y}_x} \subset \mathbb{R}^{h\times w\times C}$. Then, we aggregate the style information as:

$$\tilde{f}^t_{\tilde{y}_x} = f^{att}_{\tilde{y}_x} \odot f^t_x, \tag{5}$$

where $\odot$ means an element-wise multiplication and $\tilde{f}^t_{\tilde{y}_x}$ is the aggregated feature.

Moving forward, we adopt TPS transformation to warp $\tilde{f}^t_{\tilde{y}_x}$ to its original space as:

$$\tilde{f}_{\tilde{y}_x} = TPS(\tilde{f}^t_{\tilde{y}_x}, L_t, L_x), \tag{6}$$

where $\tilde{f}_{\tilde{y}_x}$ is aligned with the source face x. Note that, we only focus on semantic parts of makeup transfer and the other parts should be identical to the original source feature. To do this, we incorporate the source semantic mask α_x to implement feature blending [29]:

$$f_{\tilde{y}_x} = \alpha_x * \tilde{f}_{\tilde{y}_x} + (1-\alpha_x) * f_x \tag{7}$$

Discussion. Compared with the previous one-side alignment approaches [29], our canonical facial style transfer benefits from the following aspects: (1) the shorted warping distance as indicated in Eq. 2 and Eq. 3, (2) the stable semantic space with reduced semantic variance for transferring style information in such a constant facial space (illustrated in Fig. 4).

3.5 Facial Makeup Decoder (FMD)

As outlined in Section 3.4, we undertake feature interaction and transfer between the source and reference images across various feature levels $(1, 2, \cdots, M)$. In other words, we conduct canonical facial style transfer to produce $f_{\tilde{y_x}}^{1,2,\cdots,M}$ with original facial attributes but a new makeup style.

As illustrated in Fig. 3, FMD generates the m-th decoded feature $F_{\tilde{y_x}}^{m}$ by fusing the m-th level feature $f_{\tilde{y_x}}^{m}$ (produced by the CMT module in Eq. 7) and the last $(m+1)$-th decoded feature $F_{\tilde{y_x}}^{m+1}$:

$$F_{\tilde{y_x}}^{m} = \begin{cases} Conv(f_{\tilde{y_x}}^{M}), & \text{if} \quad m = M, \\ Conv(Up(F_{\tilde{y_x}}^{m+1}), f_{\tilde{y_x}}^{m}), & \text{if} \quad 1 <= m <= M-1, \end{cases} \tag{8}$$

$Up(\cdot)$ means a transposed convolutional layer that upscales the $(m+1)$ feature maps to the same size as the m-th maps. $Conv(\cdot)$ means convolutional layers. From Eq. 8, it can be seen that FMD progressively decodes the high-resolution feature map $F_{\tilde{y_x}}^{\{M,\cdots,2,1\}}$ in inverse order. And the last decoded feature $F_{\tilde{y_x}}^{1}$ is used to output the final prediction $\tilde{y_x}$ with a single convolution. Note that we set $M = 3$ in our method.

4 MT-Eval Benchmark and Cyclic Reconstruction Metrics

4.1 Benchmark

Existing methods often create test sets by randomly selecting makeup and non-makeup images from MT-Dataset [13], resulting in inconsistent standards and imbalanced difficulty levels. To address this, we design a benchmark for comprehensive evaluation across varied conditions with finer granularity.

We randomly select 8,493 source-reference image pairs from MT-Dataset to create the new MT-Eval benchmark for evaluation, using the remaining images as training data. Then, MT-Eval is categorized into three sub-testsets according to Eq. 2:

$$\mathrm{D} = \Phi(L_{src}, L_{ref}). \tag{9}$$

As shown in Fig. 5, larger spatial distances indicate greater spatial misalignments. Categorized in this manner, we collect 2,548 'small distance' pairs, 4,003 'medium distance' pairs and 1,942 'large distance' pairs, named 'MT-Eval-S', 'MT-Eval-M', 'MT-Eval-L'.

4.2 Cyclic Reconstruction Metrics for Quantitative Evaluation

Objective comparison presents a challenge due to the absence of paired images with and without makeup. And existing work usually **lacks reasonable quantitative metrics**, relying instead on subjective visual comparisons, which are insufficient for comprehensive evaluation. Leveraging shared structural content between source and its transferred result, RamGAN [26] uses SSIM for evaluation, ignoring the local structure inconsistency.

To **overall and quantitatively** evaluate the algorithm on **entire testsets**, we firstly introduce cyclic reconstruction metrics for unpaired evaluation in makeup transfer. It is inspired by some low-level methods. To this end, we perform a cyclic makeup transfer that initially maps a source image with a new makeup style y_p, and then we use the source image itself as another reference style to transfer back to the original style. The overall process can be expressed as:

$$x_{cyc}^{p} = \mathcal{G}(\mathcal{G}(x, y^{p}), x), \tag{10}$$

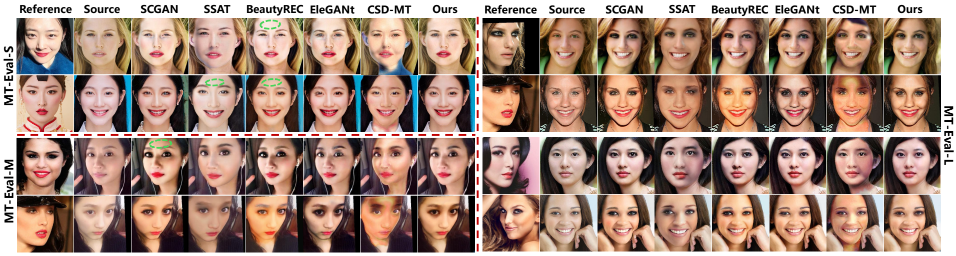

Figure 6: We present visual comparisons with SOTA makeup transfer methods on MT-Eval. The red circles highlight blurred details and artifacts, while green circles indicate inaccurate color transfers. Please zoom in for clarity. Additional visual results are provided in the supplementary materials due to space limitations.

where x^p_{cyc} is the cyclic makeup result, $\mathcal{G}$ refers to a makeup transfer model. Since x^p_{cyc} shares both style and identities with the source x, we can employ objective evaluation metrics such as PSNR and P-SSIM for accurate comparison:

$$\mathrm{PSNR}_{cyc} = \frac{1}{P}\sum_{p}^{P}(\mathrm{PSNR}(x^p_{cyc}, x)), \quad \text{P-SSIM}_{cyc} = \frac{1}{P}\sum_{p}^{P}(\text{P-SSIM}(x^p_{cyc}, x)), \tag{11}$$

We assess the cyclic reconstruction quality on the three constructed testsets, namely MT-Eval-S, MT-Eval-M, and MT-Eval-L. In each test set, we use all makeup images as y^p (P is equal to the total makeup samples in this test set.) and obtain an averaged objective result to eliminate the evaluation bias.

5 Experiment

5.1 Implementation Details and Datasets

Our project is conducted with PyTorch on an NVIDIA 3090Ti GPU. We employ the Adam optimizer with cosine annealing scheme and the learning rate starting from 2×10^{-4}. The maximum training epoch is 50 and the batch size is 1. During the inference phase, we employ the pre-trained BiSeNet [30] to automatically predict facial parsing. More training details are illustrated in our supplementary materials.

Following previous works[4, 6, 11, 13, 16, 20, 23, 26, 29], we train and evaluate our model on MT-Dataset [13], and perform a more comprehensive evaluation using our constructed benchmark MT-Eval (Sec. 4). We also test our model on the Makeup-Wild Dataset [11].

5.2 Comparison with the SOTA Methods

We compare our method with five state-of-the-art approaches SCGAN [4], SSAT [20], EleGANt [29], BeautyREC [28] and CSD-MT [21] to verify the superiority of our method. We utilize the released code and retrain these models on the same datasets as ours to conduct fair comparisons. Note that the codes and pre-trained model of RamGAN [26] are not publicly available. Moreover, for fair comparisons, we only use the official models of different methods for experiments.

Table 1: The cyclic reconstruction comparison (PSNR$_{cyc}$/P-SSIM$_{cyc}$) of different methods. The best results are highlighted in **bold**.

Testset	SSAT	BeautyREC	EleGANt	CSD-MT	Ours
MT-Eval-S	25.22/0.938	25.02/0.917	29.44/0.936	28.21/0.953	**30.76/0.953**
MT-Eval-M	25.19/0.938	24.76/0.911	29.14/0.933	27.82/0.948	**30.14/0.951**
MT-Eval-L	24.95/0.939	24.55/0.910	29.09/0.933	27.72/0.947	**29.67/0.949**

Table 2: The illustration of user study. The ratio selected as best (%).

Difficulty level	SCGAN	BeautyREC	EleGANt	CSD-MT	Ours
MT-Eval-S	5.78%	9.32%	10.93%	7.48%	**66.50%**
MT-Eval-M	8.43%	7.46%	15.88%	6.64%	**61.59%**
MT-Eval-L	6.66%	7.91%	12.03%	8.82%	**64.59%**
Makeup-Wild	11.56%	5.53%	23.12%	9.54%	**50.25%**

Qualitative Comparisons. To evaluate the models under various conditions for comprehensive comparisons, we randomly selected image pairs of three difficulty levels (MT-Eval-S, MT-Eval-M, MT-Eval-L) to form MT-Eval, as shown in Fig. 5. MT-Eval-S features frontal images with neutral expressions, MT-Eval-M includes slight profile views and mild expressions (e.g., smile), while MT-Eval-L presents significant pose and expression variations with greater spatial misalignment. Fig. 6 illustrate the qualitative comparisons on our misalignment-oriented makeup benchmark.

Quantitative Comparison. In Sec. 4, we introduce cyclic reconstruction metrics termed “PSNR$_{cyc}$" and “P-SSIM$_{cyc}$". For objective assessment, we compare our method with the other SOTA models and report the cyclic reconstruction results in Tab. 1. It is obvious that our model demonstrates the most accurate reconstruction results. Furthermore, we also provide the visual comparisons of cyclic reconstruction in supplementary materials.

User Study. We conducted a user study to evaluate the visual quality of style transfer results. We randomly selected 10 source and 10 reference images from the MT-eval and Makeup-Wild datasets, generating 100 transfer pairs per dataset. 10 predictions from each were randomly chosen and presented to 36 participants, who ranked the results based on: (1) accuracy of transferred makeup style; (2) preservation of original character identity and unrelated areas (e.g., background);, and (3) image quality of the transferred results, including any undesired artifacts and color bleeding. Tab. 2 summarizes the preference ratios across models.

5.3 Ablation Study

In this section, we conduct ablation analysis to verify the effectiveness of our designs. We provide a discussion of additional fine-grained design details in the supplementary materials.

Effectiveness of CMT. Our CMT framework minimizes spatial gaps by performing feature fusion in a canonical facial space, enabling semantic-consistent interaction. To assess this design, we train three variants with identical architectures but different alignment strategies: (1) cross-attention without alignment (‘No align’), (2) semantic warping warping to the source space (‘One-side align’), and (3) mutual warping into a canonical space (‘Canonical align’). As shown in Fig.7 (Left), the canonical align variant achieves the highest accuracy in cyclic reconstruction. Fig. 7 (Right) further demonstrates its superiority in visual quality with fewer artifacts.

Different M levels in Feature Pyramid. In Eq.1, we extract M-level pyramid features from

Testset	No align	One-side align	Canonical align
MT-Eval-S	29.83/0.925	29.96/0.927	**30.76/0.953**
MT-Eval-M	29.27/0.922	29.35/0.925	**30.14/0.951**
MT-Eval-L	28.82/0.921	28.88/0.926	**29.67/0.949**

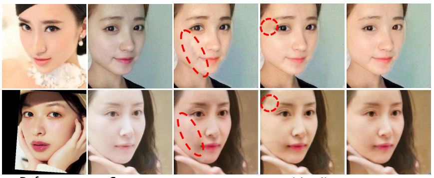

Figure 7: Evaluation of makeup transfer strategies. Left: quantitative results (PSNR$_{cyc}$↑/P-SSIM$_{cyc}$↓). Right: qualitative comparison. The 'No align' suffers from severe artifacts and color inaccuracies due to spatial misalignment, while 'One-side align' reduces these issues but still exhibits residual artifacts (marked by red circles).

Table 3: Quantitative comparison (PSNR$_{cyc}$↑/P-SSIM$_{cyc}$↓) between different M levels in pyramid feature representation.

Testset	$M = 1$	$M = 2$	$M = 3$	$M = 4$
MT-Eval-S	27.39/0.926	29.41/0.935	30.76/0.953	31.24/0.954
MT-Eval-M	27.07/0.923	29.10/0.934	30.14/0.951	30.73/0.951
MT-Eval-L	26.95/0.925	28.82/0.932	29.67/0.949	30.29/0.949

both source and reference images via the Facial Feature Extractor, and process multi-scale pairs to improve efficiency. We evaluate the effect of pyramid level M in Tab. 3. Performance improves with increasing M, with a significant gain from M=2 to M=3, but diminishing returns from M=3 to M=4. We set M=3 for an optimal balance between performance and efficiency. Qualitative results are provided in the supplementary material.

6 Conclusion

In this paper, we propose a canonical makeup transfer approach to shorten the spatial distance between source and reference image pairs, while facilitate the style transformation in a semantic consistent space. By minimizing the transformation gap, warping distortion is alleviated thus yield better information preservation and feature interaction. Furthermore, aiming at diverse conditions, we design a more comprehensive benchmark to evaluate the performance of existing methods. Extensive experiments show that our approach achieves state-of-the-art performance in transfer accuracy and artifact reduction.

Acknowledgments. This work is partially supported by Shenzhen Science and Technology Program KQTD20210811090149095 and also the Pearl River Talent Recruitment Program 2019QN01X226.

References

[1] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 40–48, 2018.

[2] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. Beautyglow: On-demand makeup transfer framework with reversible

generative network. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10042–10050, 2019.

[3] Kirsten Cowan, Ana Javornik, and Peilin Jiang. Privacy concerns when using augmented reality face filters? explaining why and when use avoidance occurs. *Psychology & Marketing*, 38(10):1799–1813, 2021.

[4] Han Deng, Chu Han, Hongmin Cai, Guoqiang Han, and Shengfeng He. Spatially-invariant style-codes controlled makeup transfer. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 6549–6557, 2021.

[5] Hajer Ghodhbani, Mohamed Neji, Imran Razzak, and Adel M Alimi. You can try without visiting: a comprehensive survey on virtually try-on outfits. *Multimedia Tools and Applications*, 81(14):19967–19998, 2022.

[6] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 10481–10490, 2019.

[7] Dong Guo and Terence Sim. Digital face makeup by example. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–79. IEEE, 2009.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[9] Ana Javornik, Yvonne Rogers, Ana Maria Moutinho, and Russell Freeman. Revealing the shopper experience of using a" magic mirror" augmented reality make-up application. In *Conference on designing interactive systems*, volume 2016, pages 871–882. Association for Computing Machinery (ACM), 2016.

[10] Ana Javornik, Yvonne Rogers, Delia Gander, and Ana Moutinho. Magicface: Stepping into character through an augmented reality mirror. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4838–4849, 2017.

[11] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5202, 2020.

[12] Chen Li, Kun Zhou, and Stephen Lin. Simulating makeup through physics-based manipulation of intrinsic image layers. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4621–4629, 2015.

[13] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 645–653, 2018.

[14] Luoqi Liu, Junliang Xing, Si Liu, Hui Xu, Xi Zhou, and Shuicheng Yan. Wow! you are so beautiful today! *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(1s):1–22, 2014.

[15] Si Liu, Xinyu Ou, Ruihe Qian, Wei Wang, and Xiaochun Cao. Makeup like a superstar: deep localized makeup transfer network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2568–2575, 2016.

[16] Thao Nguyen, Anh Tuan Tran, and Minh Hoai. Lipstick ain't enough: beyond color matching for in-the-wild makeup transfer. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 13305–13314, 2021.

[17] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

[18] Kristina Scherbaum, Tobias Ritschel, Matthias Hullin, Thorsten Thormählen, Volker Blanz, and Hans-Peter Seidel. Computer-suggested facial makeup. In *Computer Graphics Forum*, volume 30, pages 485–492. Wiley Online Library, 2011.

[19] Amila Silva, Olga Moskvyak, Alexander Long, Ravi Garg, Stephen Gould, Gil Avraham, and Anton van den Hengel. Lipat: Beyond style transfer for controllable neural simulation of lipstick using cosmetic attributes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8046–8055, January 2024.

[20] Zhaoyang Sun, Yaxiong Chen, and Shengwu Xiong. Ssat: A symmetric semantic-aware transformer network for makeup transfer and removal. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 36, pages 2325–2334, 2022.

[21] Zhaoyang Sun, Shengwu Xiong, Yaxiong Chen, and Yi Rong. Content-style decoupling for unsupervised makeup transfer without generating pseudo ground truth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7601–7610, 2024.

[22] Wai-Shun Tong, Chi-Keung Tang, Michael S Brown, and Ying-Qing Xu. Example-based cosmetic transfer. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 211–218. IEEE, 2007.

[23] Zhaoyi Wan, Haoran Chen, Jie An, Wentao Jiang, Cong Yao, and Jiebo Luo. Facial attribute transformers for precise and robust makeup transfer. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1717–1726, 2022.

[24] Shuyang Wang and Yun Fu. Face behind makeup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[25] Simon N Wood. Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1):95–114, 2003.

[26] Jianfeng Xiang, Junliang Chen, Wenshuang Liu, Xianxu Hou, and Linlin Shen. Ramgan: Region attentive morphing gan for region-level makeup transfer. In *European Conference on Computer Vision*, pages 719–735. Springer, 2022.

[27] Lin Xu, Yangzhou Du, and Yimin Zhang. An automatic framework for example-based virtual makeup. In *2013 IEEE International Conference on Image Processing*, pages 3206–3210. IEEE, 2013.

[28] Qixin Yan, Chunle Guo, Jixin Zhao, Yuekun Dai, Chen Change Loy, and Chongyi Li. Beautyrec: Robust, efficient, and component-specific makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1110, 2023.

[29] Chenyu Yang, Wanrong He, Yingqing Xu, and Yang Gao. Elegant: Exquisite and locally editable gan for makeup transfer. In *European Conference on Computer Vision*, pages 737–754. Springer, 2022.

[30] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[31] Yu-Jie Yuan, Xinyang Han, Yue He, Fang-Lue Zhang, and Lin Gao. Munerf: robust makeup transfer in neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 2024.