Received: 30 April 2024 | Accepted: 1 May 2024

DOI: 10.1002/cav.2235

SPECIAL ISSUE PAPER

WILEY

FrseGAN: Free-style editable facial makeup transfer based on GAN combined with transformer

Weifeng Xu[1] | **Pengjie Wang**[1,2] | **Xiaosong Yang**[2]

[1]School of Computer Science, Dalian Minzu University, Dalian, China

[2]National Centre for Computer Animation, Bournemouth University, Bournemouth, UK

Correspondence
Pengjie Wang, School of Computer Science, Dalian Minzu University, Dalian 116600, China.
Email: pengjiewang@qq.com

Funding information
European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No, Grant/Award Number: 900025; Natural Science Foundation of Liaoning Province, Grant/Award Number: 2023-MS-133

Abstract

Makeup in real life varies widely and is personalized, presenting a key challenge in makeup transfer. Most previous makeup transfer techniques divide the face into distinct regions for color transfer, frequently neglecting details like eyeshadow and facial contours. Given the successful advancements of Transformers in various visual tasks, we believe that this technology holds large potential in addressing pose, expression, and occlusion differences. To explore this, we propose novel pipeline which combines well-designed Convolutional Neural Network with Transformer to leverage the advantages of both networks for high-quality facial makeup transfer. This enables hierarchical extraction of both local and global facial features, facilitating the encoding of facial attributes into pyramid feature maps. Furthermore, a Low-Frequency Information Fusion Module is proposed to address the problem of large pose and expression variations which exist between the source and reference faces by extracting makeup features from the reference and adapting them to the source. Experiments demonstrate that our method produces makeup faces that are visually more detailed and realistic, yielding superior results.

KEYWORDS

generative adversarial networks, makeup transfer, transformer

1 | INTRODUCTION

Facial makeup transfer is a challenging task that involves several key issues. First, it necessitates the extraction of makeup components from the reference facial image. Second, accurately performing makeup transfer between nonaligned facial regions requires analyzing the facial structures of both faces. Finally, factors including pose, expression, lighting, and occlusion must be considered during the transfer process.

In recent years, the research of generative adversarial network (GAN) have seen rapid growth. The emergence of CycleGAN[1] has further propelled image style transfer to new heights and inspired subsequent works, such as BeautyGAN,[2] PairedCycleGAN,[3] and LADN.[4] These works can successfully transfer makeup between different faces with competitive results. And based on them, researchers propose methods for transferring of large facial poses. Representative works include PSGAN,[5] FAT,[6] and PSGAN++.[7] In the addition, researchers also explored methods for makeup transfer by fitting a 3D facial model to nonmakeup and reference images to decompose shape and texture. Representative works in this category include CPM[8] and SOGAN.[9]

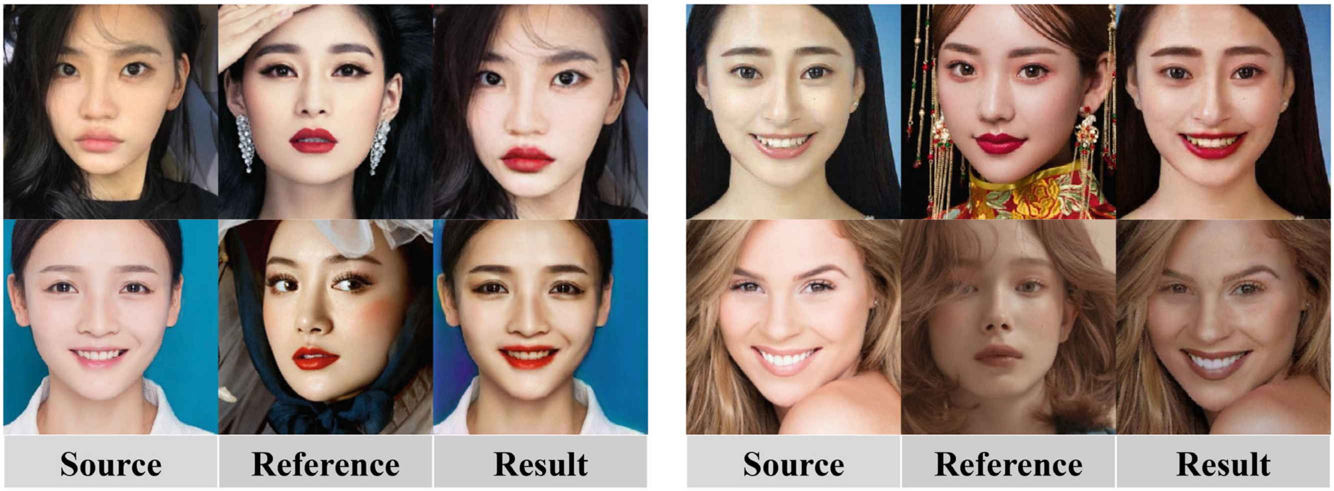

FIGURE 1 The results demonstrate the successful generation of facial images guided by the reference makeup image. The first column is the source facial image, the second column is the reference makeup image and the third column is the resulting facial image after successful makeup transfer.

Although previous works have achieved competitive results, there are still performance improvement potential by exploiting both long-distance dependency and local feature details, to address the missing facial features, blurred faces and incorrect makeup transfer faced by state-of-the-art methods. In this paper, we mainly proposed a flexible and controllable makeup transfer model by combining the Sow-Attention module from EleGANt.[10] First, unlike previous methods for face makeup transfer, our method involves a two-stage encoding process with utilizes a combination of CNN and Transformer to extract details features and well connect them to obtain a good feature representation. This representation is then fed into facial attributes encoder, and obtain pyramid-structured feature maps. Second, to tackle issues related to inconsistent head poses, we introduced a novel Low-Frequency Feature Matching (LFFM) module. This module aims to decrease computational costs while enhancing the preservation of low-frequency attributes in low-resolution feature maps. The module extracts low-frequency makeup features from the reference face and transfers them to the nonmakeup face using QKV-Attention[11] and pixel-level correlations. This design not only enables more precise extraction of various features from nonmakeup and reference faces but also has better pose robustness. Even in images with large pose differences, our method can generate good makeup transfer results. Figure 1 show that our method provides new ideas and methods for further improving makeup transfer technology.

2 | RELATED WORKS

2.1 | Vision transformer

The Transformer model was initially introduced by Google in 2017 through their paper "Attention Is All You Need."[11] Its key innovation lies in forsaking traditional CNN and RNN structures, constructing the entire network exclusively using attention mechanisms. Given its remarkable performance, fine-tuning the Transformer for downstream tasks often leads to impressive results, which naturally prompts exploration of its application to enhance computer vision tasks.

The original purpose of the Transformer was to process sequential data, rendering it unsuitable for direct application to images. To address the necessity for global relationships in computer vision tasks, researchers have used convolutional neural networks to extract 2D features from images. These features are then flattened and input into the transformer. A large achievement in this regard is DETR,[12] where Dosovitskiy et al.[13] introduced the pioneering visual Transformer (ViT), while Zhuofan Xia et al.[14] constructed an enhanced version of DAT++, which was subsequently heavily improved. Huaibo Huang et al.[15] introduced super tokens into the vision transformer. Ali Hassani et al.[16] designed the first efficient and scalable sliding-window attention mechanism for vision. Tian et al.[17] built a framework to enhance performance, based on the innovative concept of multi-resolution training. Researchers typically divide an image into smaller chunks, treating each as a word or token, similar to natural language processing. By introducing labeled tokens (class markers), the Transformer network can be directly used for image classification tasks. Anian Ruoss et al.[18] proposed a new family of positional encodings to address this challenge.

2.2 | Makeup transfer

In comparison to traditional overlay makeup techniques, makeup transfer offers high degrees of freedom, allowing users to move beyond predefined designs and autonomously select makeup styles from real model images. This significantly enriches the diversity of makeup possibilities. Moreover, makeup transfer technology can not only transfer facial makeup information but also globally migrate details such as skin tone and lighting. Currently, models based on Generative Adversarial Network (GAN) have shown promising results in this domain. BeautyGAN[2] was the first to utilize histogram loss for instance-level makeup transfer but demonstrated better transfer effects on frontal face images, with limited robustness. To address this, PSGAN[5] and PSGAN++[7] introduced an attention-based makeup deformation module to handle transformations between different head poses and facial expressions while incorporating additional features such as makeup removal. With the rise of 3D models, CPM[8] fitted a 3D face model to both nonmakeup and reference faces, using UV space maps and UV texture maps in two branches to achieve makeup transfer. In contrast, SOGAN[9] leveraged facial symmetry in UV space, performing makeup transfer solely in UV texture space. SSAT[19] employed semantic parsing maps to extract facial semantic features, which, when fused with content features, facilitated semantic alignment with reference images even in cases of large pose differences.

3 | PROPOSED METHOD

3.1 | Problem formulation

Let $X \subset R^{H\times W\times 3}$ and $Y \subset R^{H\times W\times 3}$ be the source and reference domains of facial images, respectively. It is important to note that these two image domains do not constitute paired data, meaning that the identities of individuals in the source and reference images are different. Given a nonmakeup image $x \in X$ and a makeup image $y \in Y$, the objective of the proposed method is to learn the mapping $y^x = G(x, y)$, where the generated image maintains the identity consistent with x while incorporating the makeup style of y.

3.2 | Network architecture

Figure 2 depicts the overall network architecture. Initially, both the nonmakeup face image and the reference makeup face image are fed into the Facial Attribute Reprocess Encoder to extract facial features. These features are then input into the Facial Attribute Encoder to generate a pyramid-style feature map. The high-resolution feature maps (X_H, Y_H) contain information about facial edges and details, while the low-resolution feature maps (X_L, Y_L) include information related to color and shadows. The extracted features are sent to the Makeup Transfer Module for makeup transfer. The high-frequency feature maps are aligned with the source face using a Sow-Attention module, while the Low-Frequency Information Fusion Module (LFFM) is used to align the makeup attributes of the reference face to the source face. Finally, the aligned features are multiplied with their corresponding feature maps and inputted into the Makeup Apply Decoder to produce the final makeup-transferred facial image. The functions of each module will be explained in detail below.

3.2.1 | Two-stage facial attribute encoder

To address the issue of inaccurate feature extraction for facial edges and details, we have designed the module, as depicted in Figure 2. In the facial feature extraction phase, we use a two-stage extraction, including the Facial Attribute Reprocess Encoder and the Facial Attribute Encoder. Initially, the input facial image undergoes coarse feature extraction via convolutional operations, followed by the introduction of residual connections. Between the two convolutional operations, the network's information transmission and retention capabilities are enhanced by element-wise addition of the input and output feature maps. Subsequently, the output facial feature map undergoes Patch embedding, and we introduce a

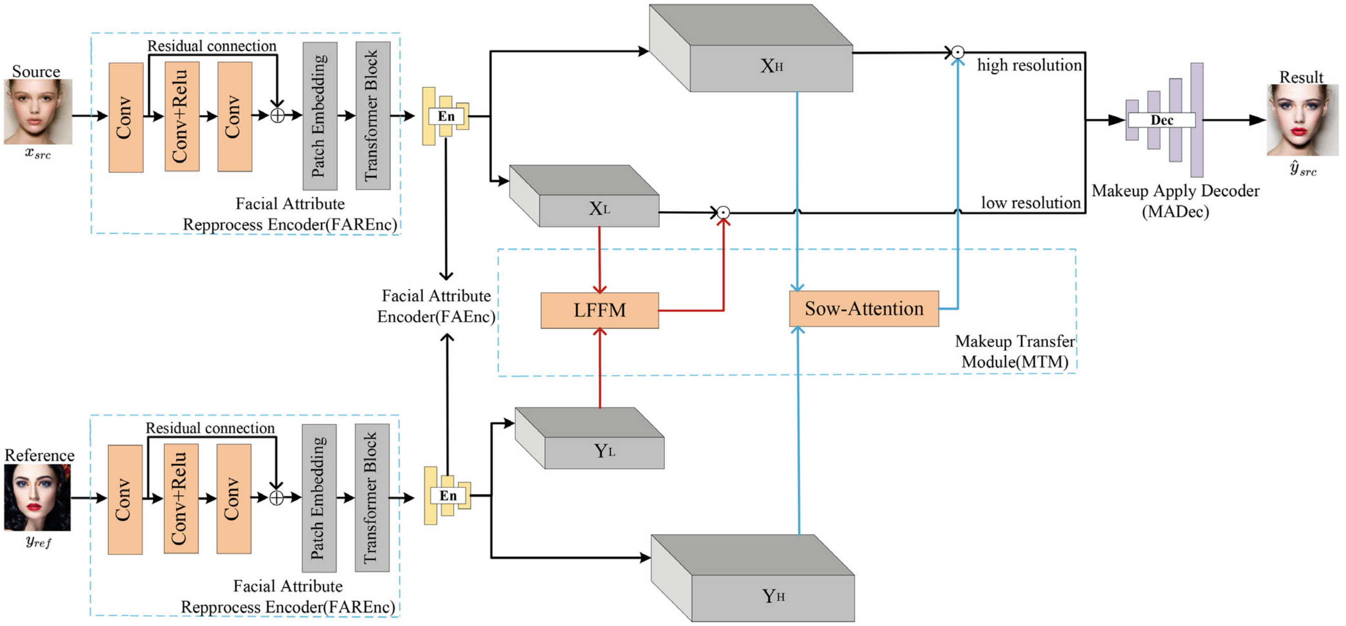

FIGURE 2 The model consists of three main parts: Two-Stage Facial Attribute Encoder(TS-FAEnc contains both FAREnc and FAEnc modules), Makeup Transfer Module (MTM) and Makeup Application Decoder (MADec).

TABLE 1 The network architecture of the face attribute encoder module (FAEnc).

Layer	Low-frequency feature map	High-frequency feature map
L_1	Conv(k:7 × 7)	Conv(k:7 × 7)
L_2	Down Sample	Down Sample
L_3	Down Sample	ResBlock×3
L_4	ResBlock×3	N/A

Transformer module inspired by Reference 20. This module preserves the Transformer's capability to model long-distance dependencies between facial features, emphasizing pixel-level interactions over channel-level interactions. Through this architecture, we aim to enhance the expressive representation of facial features.

The output feature map of the Facial Attribute Reprocess Encoder is fed into the Face Attribute Encoder module, which produces two vital feature maps: one contains high-frequency information, such as facial edges and details, while the other contains low-frequency information. The high-frequency information feature map mainly consists of micro-structures like facial edges and details. The low-frequency information feature map contains more information about the overall color, shading, and other macroscopic features of the face. Specific operations are detailed in Table 1.

3.2.2 | Makeup transfer module

Given potential differences in facial expressions and poses among faces, and recognizing that high-frequency information primarily comprises edges and details, while low-frequency information is associated with color and shadows, we use both high-resolution and low-resolution feature mappings to preserve both high-frequency and low-frequency attributes.

In this process, the Sow-Attention (Shift Overlap Window Attention Module) from Reference 10 is used to prevent excessive smoothing of high-frequency information. The module calculates attention within a shifted overlapping window, ensuring output continuity while reducing the computational cost of high-resolution input. Concurrently, two low-frequency feature maps are fed into the Low-Frequency Feature Matching (LFFM) module, the two low-frequency feature maps act as Q, K, V, and are combined with positional encoding before being fed into multi-head attention. Initially, pixel-level alignment is conducted between the reference face feature map Y_L and the source face feature map X_L.

The aligned features are then integrated, and the resulting feature map L_1 is fed into dilated convolution and depth-wise separable convolution,[21] concatenated to enhance the receptive field. Subsequently, it is input into a convolutional module, and the output feature map L_2 is generated by adding it to feature map L_1, as depicted in Figure 3, to generate a new fused feature map. This fused feature map is then multiplied with the low-frequency feature map of the original nonmakeup face image.

3.2.3 | Makeup transfer decoder module

In the final stage of our method, MADec applies the feature maps generated by the Low-Frequency Feature Matching module (LFFM) to the low-frequency information feature map X_L of the source face via element-wise multiplication. Similarly, the feature maps outputted by the Sow-Attention module are applied to the source face's high-frequency information feature map X_H through element-wise multiplication. These are inputted into the Makeup Transfer Decoder Module, ultimately producing a successfully makeup-transferred facial image.

3.3 | Loss function

3.3.1 | Adversarial loss

Adversarial loss[22] is one of the common loss functions for generative adversarial networks. Two discriminators Dx and Dy are applied to differentiate between real and generated images in domains X and Y, respectively. Therefore, the adversarial loss for discriminators and generators can be computed using the following equation, as shown in formula 2.2:

$$L_G^{adv} = -\mathbb{E}_{x\sim X,y\sim Y}\left[\log(\mathcal{D}_X(\mathcal{G}(y,x)))\right] - \mathbb{E}_{x\sim X,y\sim Y}\left[\log(\mathcal{D}_Y(\mathcal{G}(x,y)))\right], \tag{1}$$

$$\begin{aligned} L_D^{adv} = &-\mathbb{E}_{x\sim X}\left[\log \mathcal{D}_X(x)\right] - \mathbb{E}_{y\sim Y}\left[\log \mathcal{D}_Y(y)\right] - \mathbb{E}_{x\sim X,y\sim Y}\left[\log(1-\mathcal{D}_X(\mathcal{G}(y,x)))\right] \\ &- \mathbb{E}_{x\sim X,y\sim Y}\left[\log(1-\mathcal{D}_Y(\mathcal{G}(x,y)))\right], \end{aligned} \tag{2}$$

3.3.2 | Makeup loss

Makeup styles vary from person to person, and the makeup loss can have different guiding effects on their different pictures. The L_G^{make} is derived from Reference 10, and it calculates the improvement in makeup details of the generated images using the AC-PGT strategy. This loss is more stable than the histogram loss used in traditional makeup transfer within the style layer. L_G^{make} is defined as:

$$L_G^{make} = ||\mathcal{G}(x,y) - PGT(x,y)||_1 + ||\mathcal{G}(y,x) - PGT(y,x)||_1 \tag{3}$$

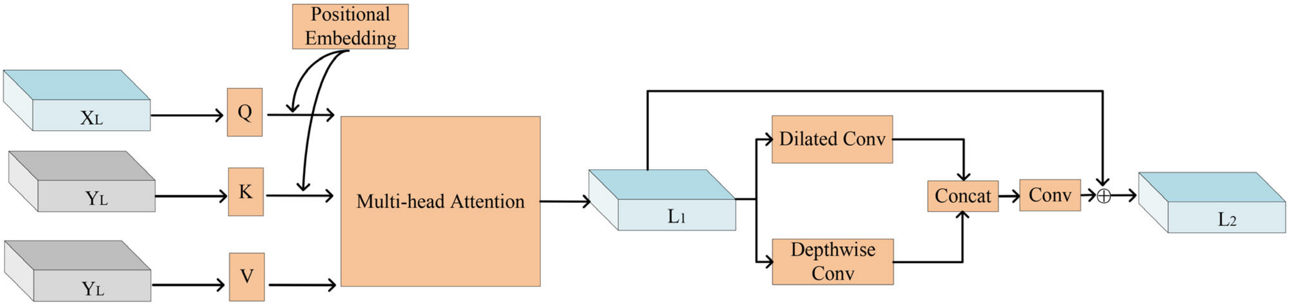

FIGURE 3 Low-Frequency Information Fusion Module (LFFM).

3.3.3 | Perceptual loss

Perceptual loss[23] can contribute to generating images that are more visually realistic. Simultaneously, to preserve personal identity information in the source image domain X during the makeup transfer process, L_G^{per} is defined as:

$$L_G^{per} = \mathbb{E}_{x\sim X,y\sim Y}\left[\|F_l(\mathcal{G}(x,y)) - F_l(x)\|_2\right] + \mathbb{E}_{x\sim X,y\sim Y}\left[\|F_l(\mathcal{G}(y,x)) - F_l(y)\|_2\right], \tag{4}$$

where $F_l(\cdot)$ is the layer 1 output of the pretrained VGG-16 model.

3.3.4 | Cycle consistency loss

As the source images lack corresponding pairs for various makeup styles, we use cycle consistency loss[1] to ensure that the generated images possess both the facial identity features of the source images and the makeup styles of the reference images. We utilize cycle consistency loss for unsupervised learning of unpaired images. The cycle consistency loss L_G^{cyc} is defined as:

$$L_G^{cyc} = \mathbb{E}_{x\sim X,y\sim Y}[\|\mathcal{G}(\mathcal{G}(x,y),x) - x\|_1] + \mathbb{E}_{x\sim X,y\sim Y}[\|\mathcal{G}(\mathcal{G}(y,x),y) - y\|_1]. \tag{5}$$

3.3.5 | Total loss

The total loss of the whole model is defined as:

$$L_{total} = \lambda_{adv}(L_G^{adv} + L_D^{adv}) + \lambda_{cyc}L_G^{cyc} + \lambda_{per}L_G^{per} + \lambda_{make}L_G^{make}, \tag{6}$$

where L_G^{adv}, L_G^{cyc}, L_G^{per} are adversarial loss, cycle consistency loss, and perceptual loss and λ_{adv}, λ_{cyc}, λ_{per}, and λ_{make} are the constant that defines the loss weights, which are set to 1, 10, 0.005, and 1.

4 | EXPERIMENTS

4.1 | Experimental details

This paper utilizes the MT dataset, which has 1115 nonmakeup images and 2719 makeup images, to train the proposed model for makeup transfer. All images are resized to 256×256 prior to training, encompassing variations in pose, expression, and background, as well as diverse makeup styles ranging from light to heavy, such as smokey, vintage, Korean-style, and American-style. The data split between training and testing sets follows the strategy outlined in BeautyGAN.[2]

4.2 | Quantitative experiments

For quantitative experiments, we randomly selected 30 and 30 nonmakeup images and makeup images from the test set to generate images respectively. In order to accurately test the quality difference of face makeup transfer results under different models, FID,[24] PSNR, and SSIM scores will be used as the indexes to measure the quality of the images, the lower the FID score proves that the quality of the generated images is better. PSNR is the most widely used image evaluation metric, but it is based on the error between the pixel points and it does not consider the visual characteristics of the human eye. So SSIM metric is added, which measures the similarity of images in terms of structure, contrast and so forth. Table 2 shows the scores of different methods for each metric, and our method gets good improvement in all three evaluation metrics. Two main reasons contribute to the effectiveness of our method: First, the Two-Stage Encoder for Facial Attributes enhances the extraction of fine-grained facial features. Second, the Low-Frequency Feature Fusion

TABLE 2 Quantitative comparison results of ablation experiments.

w/o TS-FAEnc	w/o LFFM	FID↓	PSNR ↑	SSIM ↑
		42.81	17.45	0.776
	✓	41.62	18.11	0.765
✓		40.49	18.59	0.852
✓	✓	**37.58**	**18.77**	**0.886**

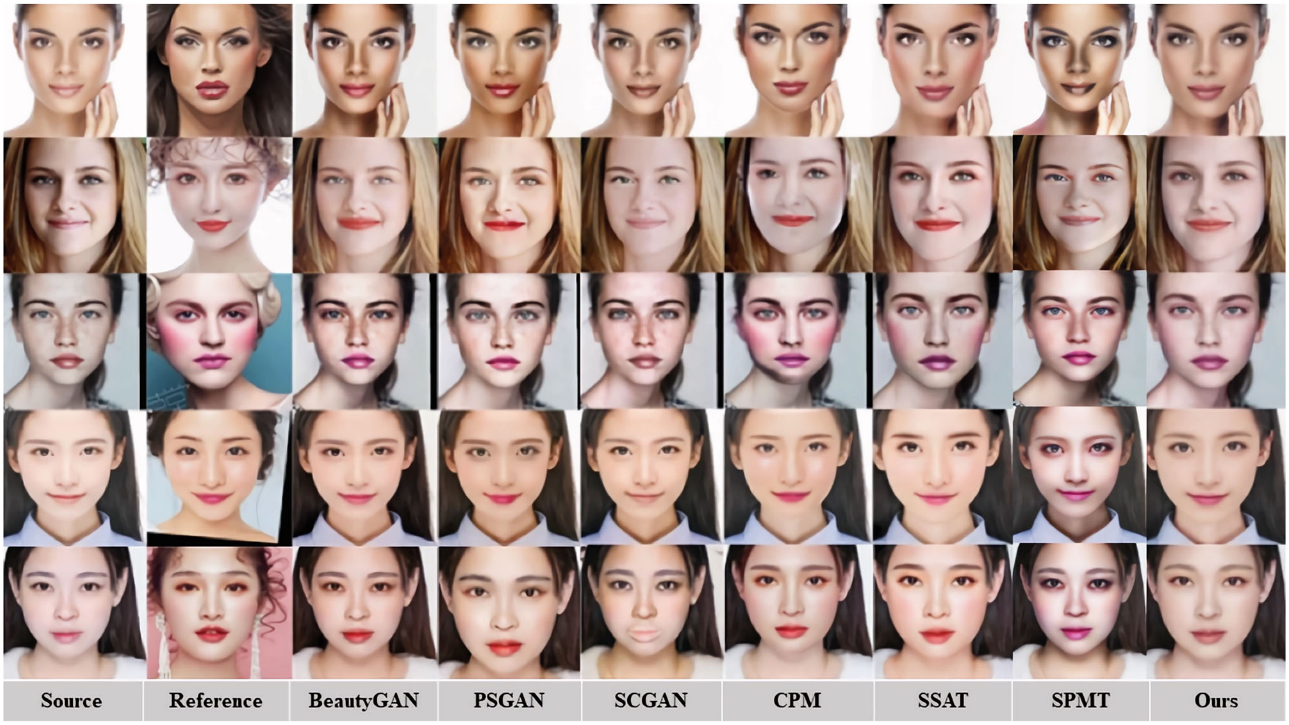

FIGURE 4 Our method is compared with BeautyGAN, PSGAN, SCGAN, CPM, SSAT, and SPMT.

Module improves the control of makeup migration details. These enhancements subsequently impact the improvement of all three evaluation metrics.

4.3 | Qualitative experiments

To begin, Figures 4 and 5 illustrate the results of personalized makeup for various expressions and styles in daily life. As shown, the methods compared may result in unnatural makeup transfer. For instance, PSGAN and SCGAN, although capable of producing visually acceptable results, suffer from inaccurate color transfer, such as blurry lip edges and partial pseudo-shadows in the eye area. CPM exhibits noticeable pseudo-shadows along facial edges in the transfer results. Additionally, SCGAN and SSAT struggle to preserve nonmakeup areas, such as faded backgrounds, and even compromise the identity of the source image, such as in the eyes. Our method solves these problems that exist in other methods of makeup transfer. This means that our method can take into account both high and low frequency information of the face, which is due to our idea of using two innovative modules for different frequency control of the makeup transfer generated image, which greatly reduces the problem of incorrect color transfer from different regions during makeup transfer, and improves the retention of extraneous makeup regions at the same time. In contrast, our method not only effectively transfers makeup styles but also excels in preserving identity and nonmakeup areas compared to existing methods. Overall, our proposed method yields more realistic results.

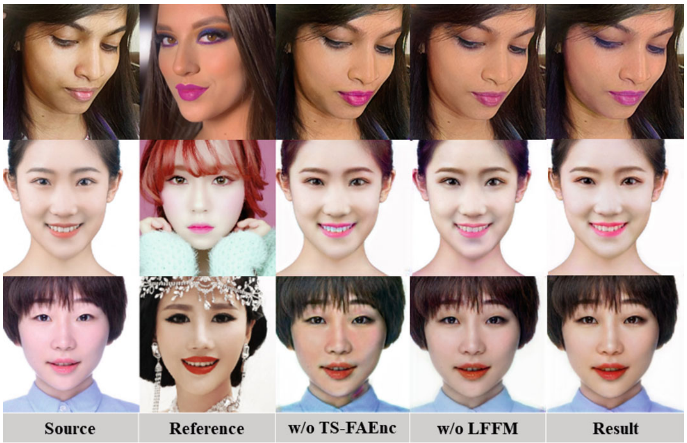

FIGURE 5 The ablation experiment results with the first column displaying the input source facial image, the second column showing the reference makeup facial image, the third column presenting the generated effect without the Two-Stage Encoder for Facial Attributes. The fourth column illustrates the result without the low-frequency feature fusion module. The fifth column demonstrates that the results generated based on the innovation proposed in this paper do not have these problems.

4.4 | Ablation experiments

4.4.1 | Impact of two-stage encoder for facial attributes

To assess the impact of Two-Stage Encoder for Facial Attributes on the effectiveness of transferring makeup styles from reference images to bare-faced images, this study removed the module and trained the model using only the facial attribute encoder. The training results fails to completely transfer the makeup style from the reference image to the nonmakeup image, as shown in Figure 5. This module fails to capture some finer details, leading to color discrepancies in areas such as eyeshadow and teeth. Compared to the results without using Two-Stage Encoder for Facial Attributes, the patterns' outlines are more accurate and clear when Two-Stage Encoder for Facial Attributes is used and the facial pattern colors closely resemble those in the reference makeup images. The quantitative test results after removing the TS-FAEnc module are shown in Table 2, confirming that the generated results are less similar to pseudo-paired images, indicating higher accuracy in image translation.

4.4.2 | Impact of the low-frequency feature fusion module

To validate the necessity of the proposed low-frequency information fusion module, experimental results are presented in Figure 5. First in the generated images without LFFM (first row, fourth column), the eyeshadow is not successfully transferred, but with LFFM the eyeshadow is successfully and clearly visible, and the lipstick color is pronounced. Second, in the generated images without LFFM (second row, fourth column), the teeth color remains contaminated and incorrect colors are present in the shadow under the lips and chin. In the final generated images with LFFM, the teeth color is clean and uniform. Lastly, in the generated images without LFFM (third row, fourth column), the facial skin color does not match the reference makeup face. In the final generated images with LFFM, the facial skin color matches the reference makeup face, showing a clear distinction. In addition to the subjective visual results, this study quantitatively compares the generated results, as shown in Table 3. It can be observed that all three metrics are affected, indicating the effectiveness of this module.

4.5 | Additional visual effects

This study further conducts makeup transfer experiments on images with large variations in facial poses, as depicted in Figures 6 and 7. For images with considerable changes in facial poses, the proposed method still achieves satisfactory

TABLE 3 Our method is compared with BeautyGAN, PSGAN, SCGAN, CPM, SSAT, and SPMT on three evaluation metrics.

	FID↓	PSNR ↑	SSIM ↑
BeautyGAN[2]	55.92	14.74	0.72
PSGAN[5]	56.50	15.72	0.73
SCGAN[25]	82.55	13.97	0.48
CPM[8]	47.17	13.55	0.69
SSAT[19]	44.02	17.56	0.81
SPMT[26]	42.36	17.78	0.83
Ours	**37.58**	**18.77**	**0.88**

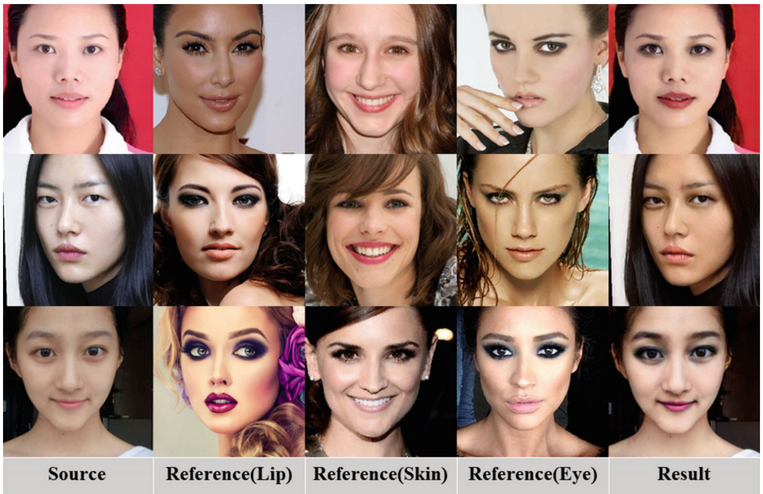

FIGURE 6 The results of multiple reference image blending and transfer.

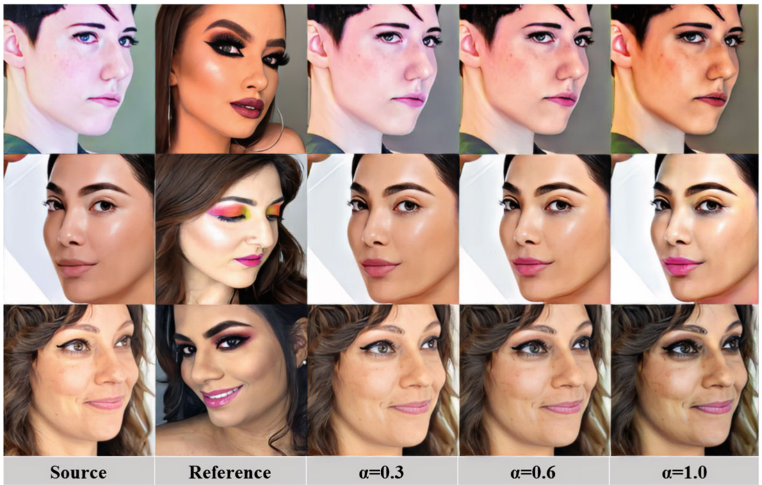

FIGURE 7 The results of the style interpolation experiment.

makeup transfer results. This is attributed to the accurate establishment of correspondence between two faces during the fusion of low-frequency and high-frequency information, naturally preserving expressions and poses between source and reference images.

The makeup transfer module outputs high-frequency and low-frequency information feature maps, corresponding spatially to the nonmakeup facial regions. This is achieved by combining the transferred high-frequency and low-frequency information feature maps with source feature maps using element-wise multiplication, allowing for controllability through interpolation of these makeup feature maps. First, the mask type for generating facial parts is determined, including full (entire face), lip (lip region), skin (skin area), and eye (eye region). Then based on the mask type corresponding logic is selected to generate the appropriate mask. The coefficient α represents the color intensity of the generated mask, ranging from [0, 1], indicating the color intensity coefficient of the mask, gradually increasing from 0 to 1. By weighting the makeup feature tensors from two reference makeup images using the coefficient α, a new makeup tensor is obtained. As showed in Figure 7, this model can gradual improvement of makeup transfer effects is observed in the resulting images of successful makeup adjustments with different coefficients α set to 0.3, 0.6, and 1.0 and achieve controllable blending of multiple makeup styles by combining mask types for different facial regions, allowing simultaneous fusion of nonmakeup images with reference makeup styles one, two, and three.

5 | CONCLUSION

We propose a novel facial makeup transfer method that excels in transferring details. Our method simultaneously leverages high-resolution and low-resolution feature to retain both high-frequency and low-frequency attributes. To establish feature correspondences between different faces and reduce computational complexity, we introduce an innovative LFFM module. Extensive experiments demonstrate the robustness of our proposed method in handling variations in facial poses. Nonetheless, our method fails in situation when dealing with intricate facial paintings, resulting in suboptimal results. Therefore, we will focus in primarily on transferring complex makeup styles.

ACKNOWLEDGMENTS

Our work is supported by European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 900025; Natural Science Foundation of Liaoning Province, Grant/Award Number: 2023-MS-133.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Weifeng Xu https://orcid.org/0009-0002-7960-0377

REFERENCES

1. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision; 2017. p. 2223–32.
2. Li T, Qian R, Dong C, Liu S, Yan Q, Zhu W, et al. BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network. 2018 ACM Multimedia Conference; 2018.
3. Chang H, Lu J, Yu F, Finkelstein A. PairedCycleGAN: Asymmetric style transfer for applying and removing makeup. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018.
4. Gu Q, Wang G, Chiu MT, Tai YW, Tang CK. LADN: Local adversarial disentangling network for facial makeup and de-makeup. 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019.
5. Jiang W, Liu S, Gao C, Cao J, Yan S. PSGAN: Pose and expression robust spatial-aware GAN for customizable makeup transfer. 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2020.
6. Wan Z, Chen H, An J, Jiang W, Yao C, Luo J. Facial attribute transformers for precise and robust makeup transfer. Proceedings of the IEEE/CVF winter conference on applications of computer vision; 2022. p. 1717–26.
7. Liu S, Jiang W, Gao C, He R, Yan S. PSGAN++: Robust detail-preserving makeup transfer and removal. IEEE Trans Pattern Anal Mach Intell. 2021a;PP(99):1.
8. Nguyen T, Tran AT, Hoai M. Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer. Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition; 2021. p. 13305–14.

9. Lyu Y, Dong J, Peng B, Wang W, Tan T. SOGAN: 3D-aware shadow and occlusion robust GAN for makeup transfer. Proceedings of the 29th ACM International conference on multimedia; 2021. p. 3601–9.
10. Yang C, He W, Xu Y, Gao Y. Elegant: Exquisite and locally editable gan for makeup transfer. European Conference on Computer Vision. Springer; 2022. p. 737–54.
11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Proces Syst. 2017;30.
12. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. European conference on computer vision. Springer; 2020. p. 213–29.
13. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint, arXiv:2010.11929. 2020.
14. Xia Z, Pan X, Song S, Li LE, Huang G. DAT++: Spatially dynamic vision transformer with deformable attention. arXiv preprint, arXiv:2309.01430. 2023.
15. Huang H, Zhou X, Cao J, He R, Tan T. Vision transformer with super token sampling. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023;22690–22699.
16. Hassani A, Walton S, Li J, Li S, Shi H. Neighborhood attention transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 6185–94.
17. Tian R, Wu Z, Dai Q, Hu H, Qiao Y, Jiang YG. Resformer: Scaling vits with multi-resolution training. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 22721–31.
18. Ruoss A, Delétang G, Genewein T, Grau-Moya J. Randomized positional encodings boost length generalization of transformers. arXiv preprint, arXiv:2305.16843. 2023.
19. Sun Z, Chen Y, Xiong S. Ssat: A symmetric semantic-aware transformer network for makeup transfer and removal. Proceedings of the AAAI Conference on artificial intelligence. Volume 36; 2022. p. 2325–34.
20. Qiu Y, Zhang K, Wang C, Luo W, Li H, Jin Z. MB-TaylorFormer: Multi-branch efficient transformer expanded by Taylor formula for image dehazing. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 12802–13.
21. Chollet F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017;1251–1258. 2016.
22. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Adv Neural Inf Proces Syst. 2014;27.
23. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Berlin, Germany: Springer International Publishing; 2016. p. 694–711.
24. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. 2017.
25. Deng H, Han C, Cai H, Han G, He S. Spatially-invariant style-codes controlled makeup transfer. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021a. p. 6549–57.
26. Zhu M, Yi Y, Wang N, Wang X, Gao X. Semi-parametric makeup transfer via semantic-aware correspondence. arXiv preprint, arXiv:2203.02286. 2022.

AUTHOR BIOGRAPHIES

Weifeng Xu is currently a postgraduate student with School of Computer Science, Dalian Minzu University, Dalian, China. Her research interests include computer vision.

Pengjie Wang is currently a professor with School of Computer Science, Dalian Minzu University, Dalian, China. His research interests include computer vision and computer graphics.

Xiaosong Yang is currently a Professor, Deputy Head of Department, Programme Leader of MSc AIM at the National Centre for Computer Animation, Bournemouth University. He has over 30 years' experience of research, education and professional practice in computer animation, machine learning, data mining, digital health, virtual reality and surgery simulation.

How to cite this article: Xu W, Wang P, Yang X. FrseGAN: Free-style editable facial makeup transfer based on GAN combined with transformer. Comput Anim Virtual Worlds. 2024;35(3):e2235. https://doi.org/10.1002/cav.2235