Content-Style Decoupling for Unsupervised Makeup Transfer without Generating Pseudo Ground Truth

Supplementary Material

1. Network Structure

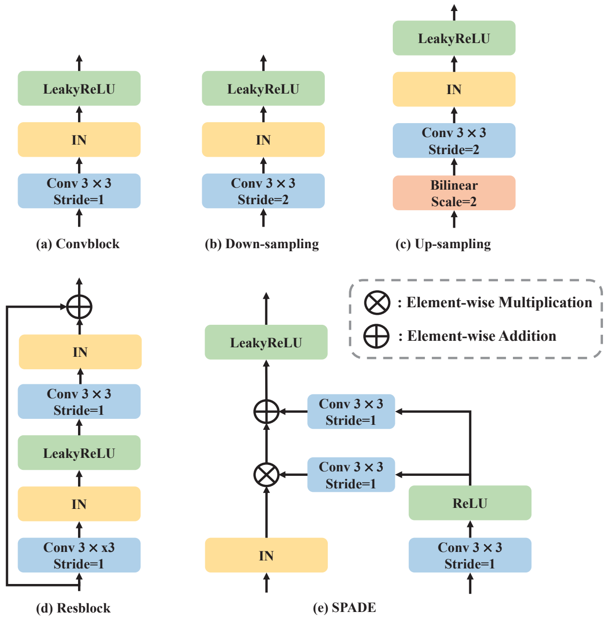
We design 5 basic network blocks to construct the generator $\mathcal{G}$ in our CSD-MT model, including Convblock, Down-sampling block, Up-sampling block, Resblock, and SPADE, whose structures are shown in Figure 1. Based on these blocks, Figure 2 illustrates the architectures of the semantic correspondence module and the makeup rendering module, where the shape of each intermediate feature map is also presented. Note that, before feeding the input images into the semantic correspondence module, we concatenate them with their corresponding face parsing [11] maps to enhance the local semantic information of different facial parts (in our implementation, the face parsing maps of 10 semantic categories are utilized). To align with the target distribution, the proposed CSD-MT model adopts the same multi-scale discriminator $\mathcal{D}$ as in [9], which consists of 3 scale-specific discriminators trained at 3 different image scales with an identical architecture.

2. Training Details

During the model training, each input image is manually resized to 256 × 256 pixels. In our color contrastive loss, four negative samples are generated for each transferred image, i.e., $N = 4$ in Eq. (8). And the feature maps from the `relu_1_2`, `relu_2_2` layers of the pre-trained VGG19 model are used for calculating gram matrices (see Eq. (9)). For the hyper-parameters, we set $\tau = 100$ in Eq. (2), $\alpha = 0.1$ in Eq. (4), and $\lambda_{trans} = 1$, $\lambda_{cycle} = 10$, $\lambda_{adv} = 1$, $\lambda_{aug} = 10$, $\lambda_{cts} = 1$ in Eq. (10). We use the Adam [6] optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for model training, the maximum number of training iterations is 500,000, the learning rate is 0.0002, and the batch size is 1.

3. Parameter Size and Inference Speed

In addition to the makeup transfer performance, we also compare the parameter size and inference time of CSD-MT with those of the competing methods. For a fair comparison, all the experiments are conducted on a single NVIDIA GTX 1660Ti GPU with 6GB RAM. From the results in Table 1, it can be seen that our CSD-MT model has the least number of parameters (6.94 M) and achieves the fastest inference speed (only 0.017 seconds for processing a pair of input images with a resolution of 256 × 256 pixels), which surpasses other benchmark methods by a large margin. This indicates the efficiency of our CSD-MT method.



(a) Convblock (b) Down-sampling (c) Up-sampling

(d) Resblock (e) SPADE

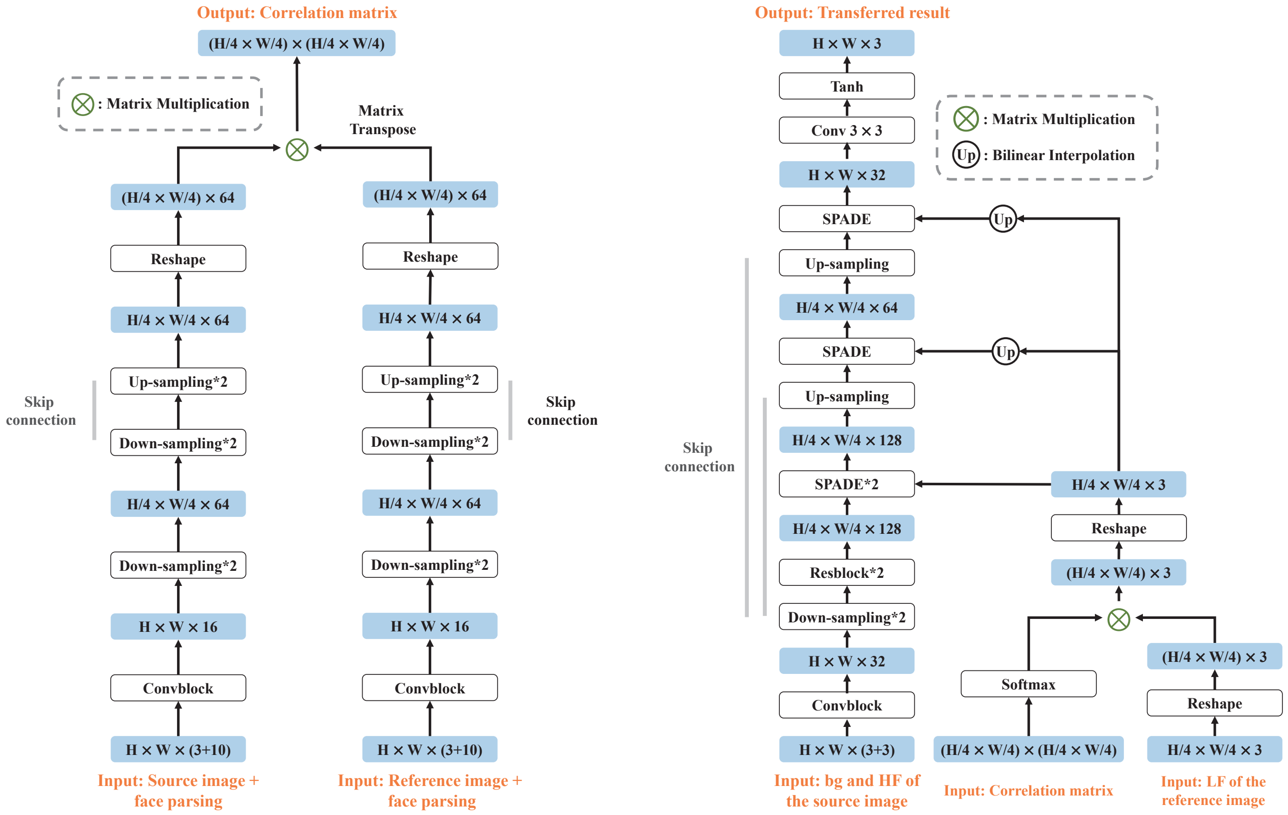
Figure 1. Basic network blocks used in the proposed CSD-MT model. Here, ”IN” denotes an instance normalization layer.

4. Trade-off between Content and Makeup

By minimizing the transfer loss L_{trans} (see Eq. (4) in the main text), CSD-MT simultaneously preserves the content details in the source image (L_{cont}) and transfers the makeup information of the reference face (L_{makeup}). There is a trade-off between these two objectives, which is balanced by the importance parameter α. To investi-

Methods	Parameters (M)	Inference Time (s)
BeautyGAN [7]	9.42	0.039
PSGAN [4]	12.62	0.218
SCGAN [2]	15.30	0.321
SpMT [14]	333.67	0.834
LADN [3]	27.00	0.032
SSAT [8]	10.48	0.110
EleGANt [10]	10.27	0.148
CSD-MT (ours)	**6.94**	**0.017**

Table 1. Comparisons of the parameter size and inference speed of CSD-MT and other methods. The number of parameters (M) and inference time (seconds) are calculated for different models when processing a pair of input images with a size of 256 × 256 pixels.



(a) Semantic Correspondence Module

(b) Makeup Rendering Module

Figure 2. The architecture of the semantic correspondence module and makeup rendering module in CSD-MT. *n indicates a stack of n blocks.

Parameter	Self-Aug PSNR/SSIM	
	Crop	Rotate
$\alpha = 0.0$	23.77/0.830	22.12/0.791
$\alpha = 0.1$	**27.28/0.920**	**26.68/0.915**
$\alpha = 0.5$	24.91/0.908	24.62/0.906

Table 2. Quantitative comparison of CSD-MT models trained with different α on the MT dataset.

gate the effect of this parameter, we compare the performance of CSD-MT models trained with different values of α (varying in $\{0.0, 0.1, 0.5\}$). Both quantitative and qualitative comparisons are conducted. As shown in Table 2, the proposed CSD-MT method achieves the best self-augmented PSNR/SSIM results when $\alpha = 0.1$ (27.28/0.920 and 26.68/0.915 on "Crop" and "Rotate" scenarios, respectively). Such phenomenon can also be found in Figure 3. When $\alpha = 0.0$, the content objective L_{cont} is removed from L_{trans}, so the trained model fails to retain the content information in the source images and generates unrealistic results. When the value of α increases to 0.5, L_{cont}

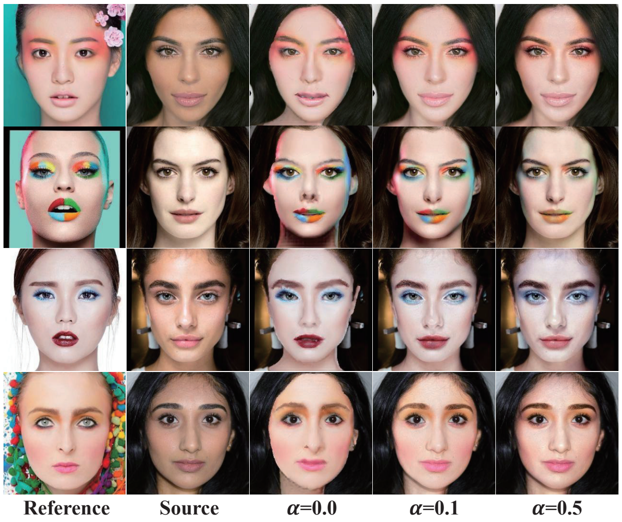


Figure 3. Qualitative comparison of CSD-MT models trained with different α. $\alpha = 0.1$ leads to the best transferred results.

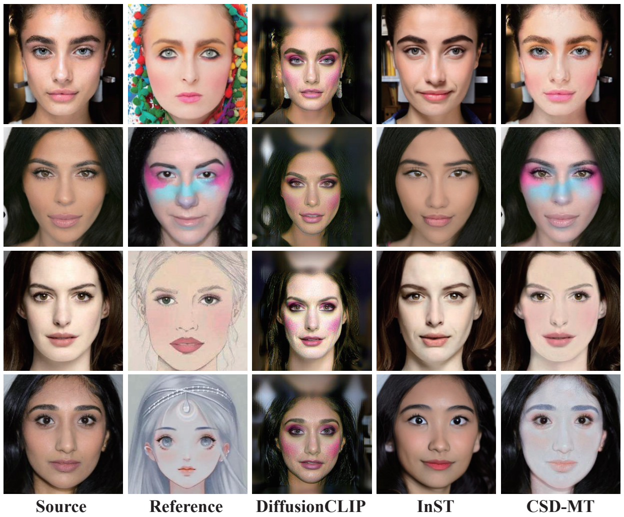


Figure 4. Qualitative comparison with diffusion models, including DiffusionCLIP [5] and InST [13].

dominates the transfer loss L_{trans} and reduces the relative importance of L_{makeup}. As a result, the makeup styles of the reference faces, especially lipstick and powder blush, cannot be faithfully transferred.

5. Comparison with Diffusion Models

Recently, powerful diffusion models have been widely studied and become mainstream approaches for solving various image generation tasks. Therefore, we would also like to compare our CSD-MT method with diffusion models. Considering that there is currently no diffusion model specifically designed for the makeup transfer task, a text-guided generative diffusion model DiffusionCLIP [5] and a style transfer diffusion model InST [13] are chosen as the benchmark methods. For DiffusionCLIP, since it is difficult accurately describe a specific makeup style in text, we use the prompt ”people with makeup” as in [5] to produce the final transferred results. From Figure 4, it can be seen that DiffusionCLIP usually introduces incorrect makeup information in the final outputs, since its generation process is mainly based on the text prompt instead of the reference image. As a style transfer method, InST not only fails to distill makeup styles from the reference image but also alters the content details of the source image. CSD-MT outperforms these two diffusion model based methods, again demonstrating its effectiveness and superiority.

6. Makeup Control

6.1. Makeup Removal

Similar to [3, 8, 12], by taking makeup images as the source inputs and non-makeup faces as the reference images, CSD-MT can also generate multiple makeup removal results, as displayed in Figure 5.

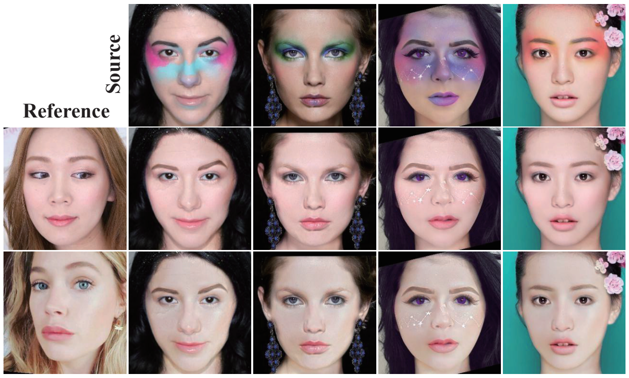


Figure 5. The makeup removal results generated by CSD-MT.

6.2. Global Makeup Interpolation

In our proposed CSD-MT method, the makeup information are decoupled from the input images through frequency decomposition. This allows us to interpolate the makeup styles between two different reference faces by linearly fusing their low-frequency (LF) components, as follows:

$$\begin{aligned} \hat{y}_l^{g\_inter} &= (1-\beta)\hat{y}_l^1 + \beta\hat{y}_l^2, \\ \hat{x}^{g\_inter} &= G_{mr}([x_{bg}, x_h], \hat{y}_l^{g\_inter}). \end{aligned} \tag{1}$$

Here $\hat{y}_l^1$ and $\hat{y}_l^2$ are deformed LF components of two different reference images, respectively. By adjusting the value of β from 0 to 1, CSD-MT can generate a series of transferred results. Their makeup styles will gradually change from that of one reference image y^1 to that of the other y^2. Moreover, by assigning the source image as y^1, we can control the degree of makeup transfer for a single reference input y^2. The global makeup interpolation results are shown in Figure 6.

6.3. Local Makeup Interpolation

In CSD-MT, the LF component of the reference image is deformed through the correlation matrix M, so that it can be semantically aligned with the source image. Such spatial alignment enables CSD-MT to implement the makeup interpolation within different local facial areas, which can be formulated as follows:

$$\begin{aligned} \hat{y}_l^{l\_inter} &= ((1-\beta)\hat{y}_l^1 + \beta\hat{y}_l^2) \otimes Mask_x^{area} \\ &+ \hat{x}_l \otimes (1 - Mask_x^{area}), \\ \hat{x}^{l\_inter} &= G_{mr}([x_{bg}, x_h], \hat{y}_l^{l\_inter}). \end{aligned} \tag{2}$$

where $\otimes$ denotes the Hadamard product. $Mask_x^{area}$ is a binary mask of the source image x, indicating the local areas to be makeup, which can be obtained by face parsing. Figure 7 visualizes the local makeup interpolation results

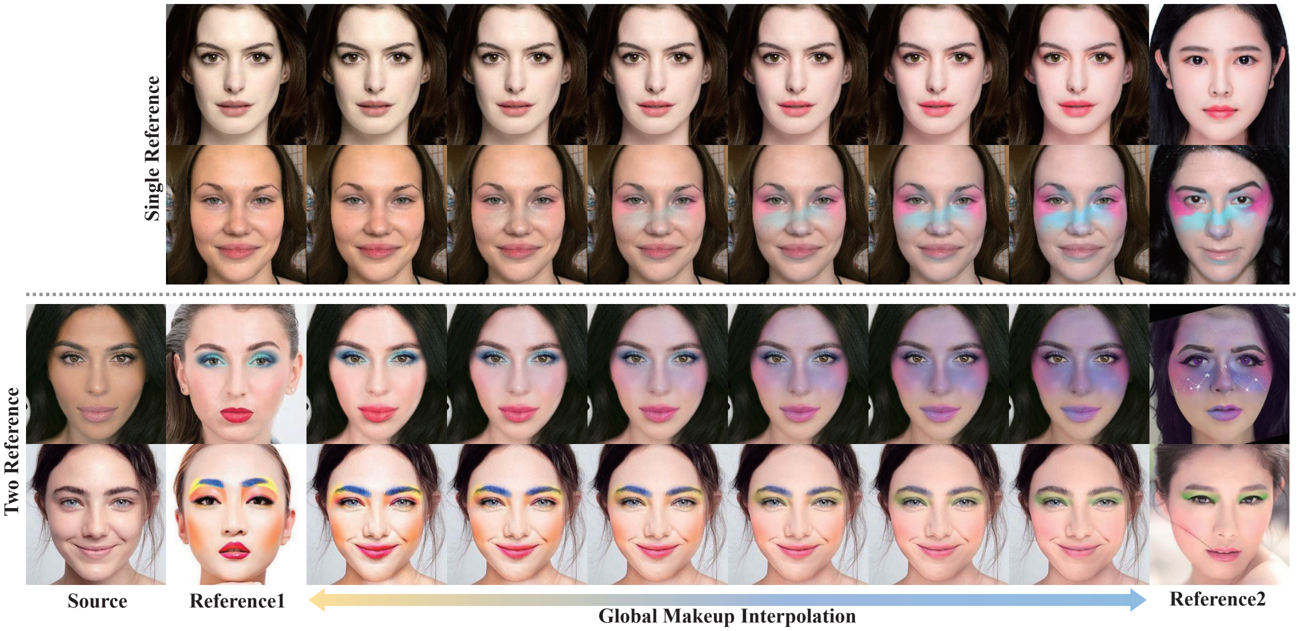


Figure 6. The illustration of global makeup interpolation. The first two rows are the result of a single reference image, the last two rows are the result of two reference images.

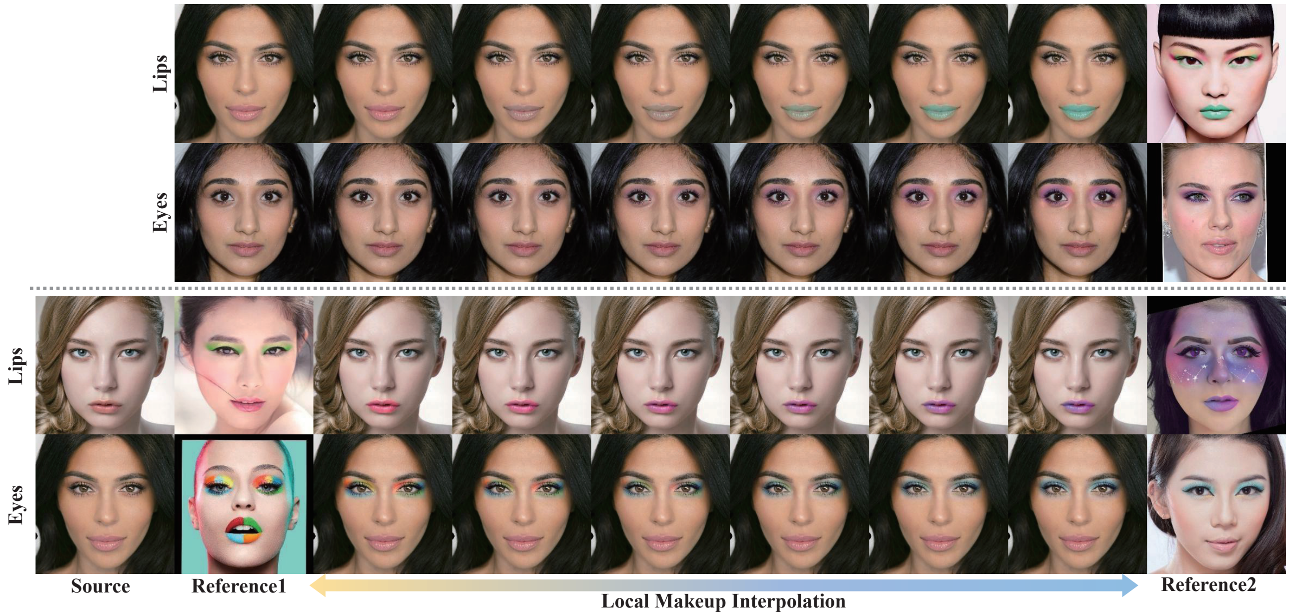


Figure 7. The illustration of local makeup interpolation. The odd rows are lipstick control, the even rows are eye shadow control.

within the areas around the lips and eyes, respectively, i.e., $area \in \{lip, eye\}$ for $Mask_x^{area}$. Similarly, we can also control the local makeup transfer degree of a single reference image by replacing the other reference input with the source image, as shown in the first two rows of Figure 7.

Preserving Skin Tone. Similar to previous approaches [1–4, 7, 8, 10, 14], CSD-MT assumes that the foundations and other cosmetics have already covered the original skin tone. Therefore, the skin color of the reference face is considered as a part of its makeup styles and is faithfully transferred to the final generated result, which may corrupt the content information in the source image. To alleviate this problem, we can perform the above-mentioned local makeup interpolation operation in the face region of the source image to

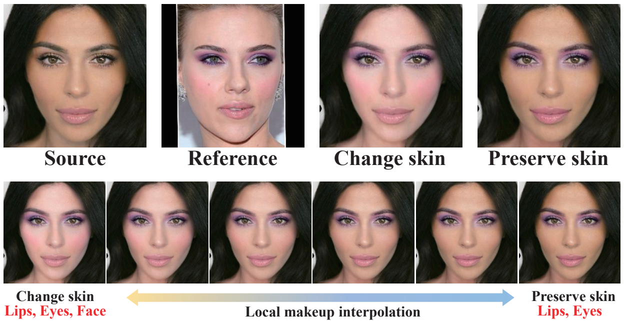


Figure 8. By default, our method CSD-MT transfers makeup to change the skin tone. Optionally, the local makeup transfer operation can preserve the original skin tone, and the local makeup interpolation can smoothly generate intermediate results.

preserve its skin tone. This procedure can be formulated as:

$$\begin{aligned} \hat{y}_l^{l\_skin} =& ((1-\beta)\hat{x}_l + \beta\hat{y}_l^2) \otimes Mask_x^{face} \\ &+ \hat{y}_l^2 \otimes (1 - Mask_x^{face}), \\ \hat{x}^{l\_skin} =& G_{mr}([x_{bg}, x_h], \hat{y}_l^{l\_skin}). \end{aligned} \tag{3}$$

Here, $\hat{x}^{l\_skin}$ realizes the local makeup interpolation between the source image x and the reference image y^2 within the face region in x, which is indicated by the mask $Mask_x^{face}$. The interpolation results are visualized in Figure 8. When $\beta = 0$, $\hat{x}^{l\_skin}$ will not change the skin tone of x. And when $\beta = 1$, Eq. (3) degenerates to the standard makeup transfer process in CSD-MT, which will distill the makeup information (including the skin tone) from y^2 to x.

6.4. Partial Makeup Transfer

In addition, CSD-MT can integrate local makeup styles from different reference images for partial makeup transfer.

$$\begin{aligned} \hat{y}_l^{part} =& \hat{y}_l^1 \otimes Mask_x^{lip} + \hat{y}_l^2 \otimes Mask_x^{eye} \\ &+ \hat{y}_l^3 \otimes Mask_x^{face}, \\ \hat{x}^{part} =& G_{mr}([x_{bg}, x_h], \hat{y}_l^{part}). \end{aligned} \tag{4}$$

where $Mask_x^{lip}$, $Mask_x^{eye}$, $Mask_x^{face}$ are the lip, eye and face masks of the source image x. The results of partial makeup transfer are shown in Figure 9.

6.5. Makeup Editing

CSD-MT also allows users to create their own customized makeup looks by editing the reference image. This editing process is simple and intuitive, the users only need to apply their preferred colors to any local area of the reference face. After that, our CSD-MT model is employed to transfer these user-edited makeup styles to the source images. As shown in Figure 10, CSD-MT generates better transferred results compared to other state-of-the-art methods.

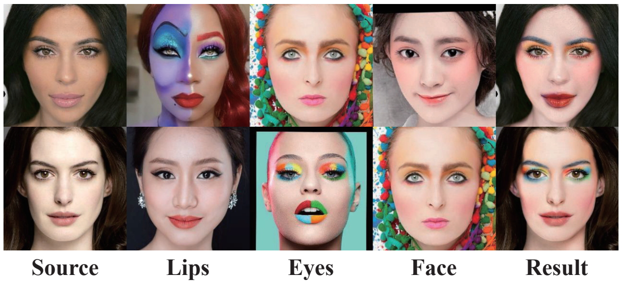


Figure 9. The results of partial makeup transfer. The results integrate the lips style from the second column, the eyes style from the third column, and the face style from the fourth column.

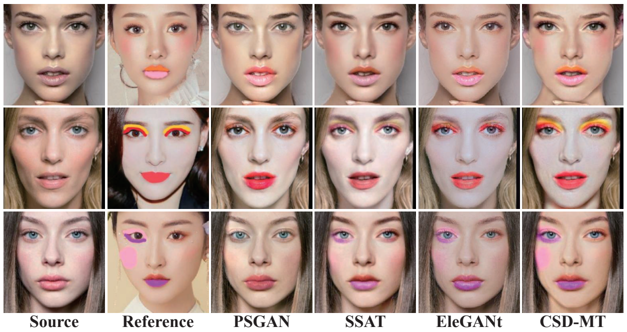


Figure 10. Comparison of makeup editing with different methods.

7. More Results

Figure 11, Figure 12, and Figure 13 show more qualitative comparisons between CSD-MT and state-of-the-art methods under simple, complex, and extreme makeup styles, respectively. More makeup transfer results of CSD-MT are shown in Figure 14 and Figure 15. Additionally, the robustness in various complex scenarios is demonstrated in Figure 16, the generalization ability to unseen makeup styles is shown in Figure 17, and the control ability over makeup editing is illustrated in Figure 18.

8. The Limitation

In CSD-MT, we assume that the high-frequency (HF) component is more closely associated with the content details of face images. With this assumption, CSD-MT preserves content details by maximizing the consistency of high-frequency information between the source image and the transferred result. As a result, certain boundaries (HF information) of some extreme makeup are treated as content details rather than makeup style in CSD-MT. Please refer to the makeup removal result in the fourth column of Figure 5. At the same time, our CSD-MT is ineffective in accurately rendering the boundaries of some extreme makeup styles, as shown in Figure 19. In the future, our research will primarily focus on finding solutions to this problem.

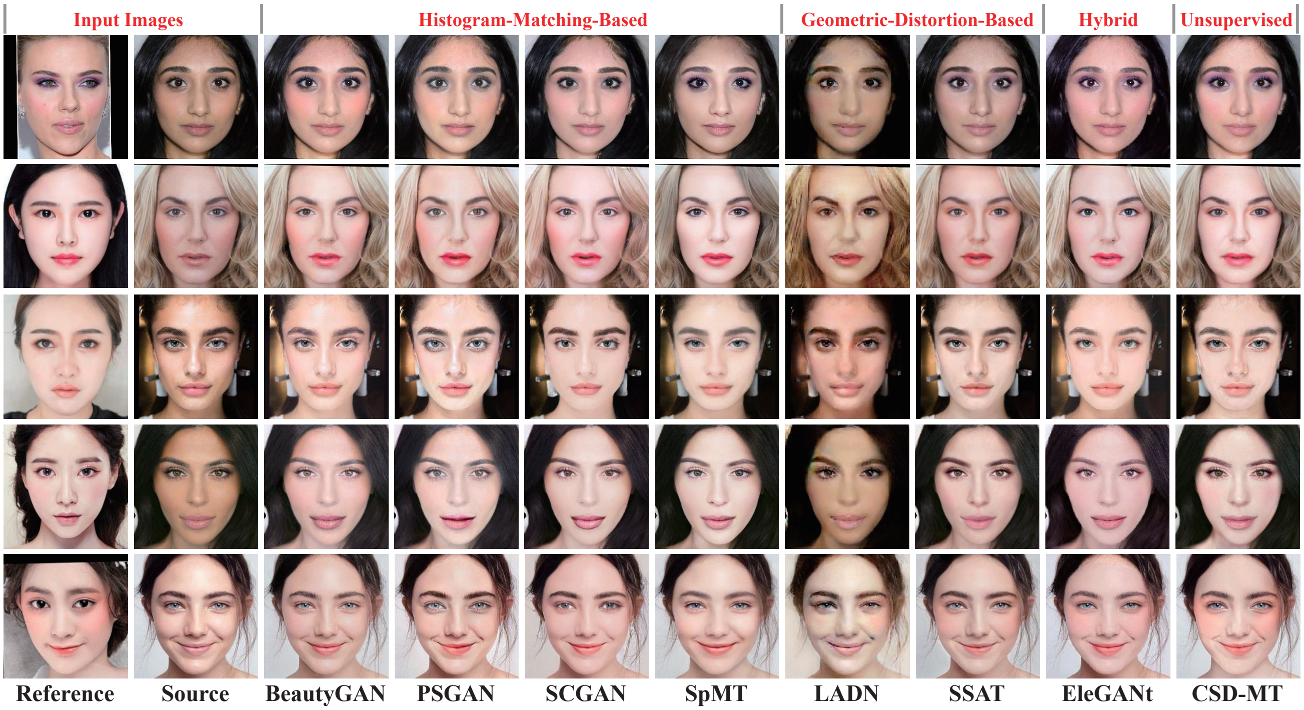


Figure 11. More qualitative comparisons between CSD-MT and state-of-the-art methods **under simple makeup styles**.

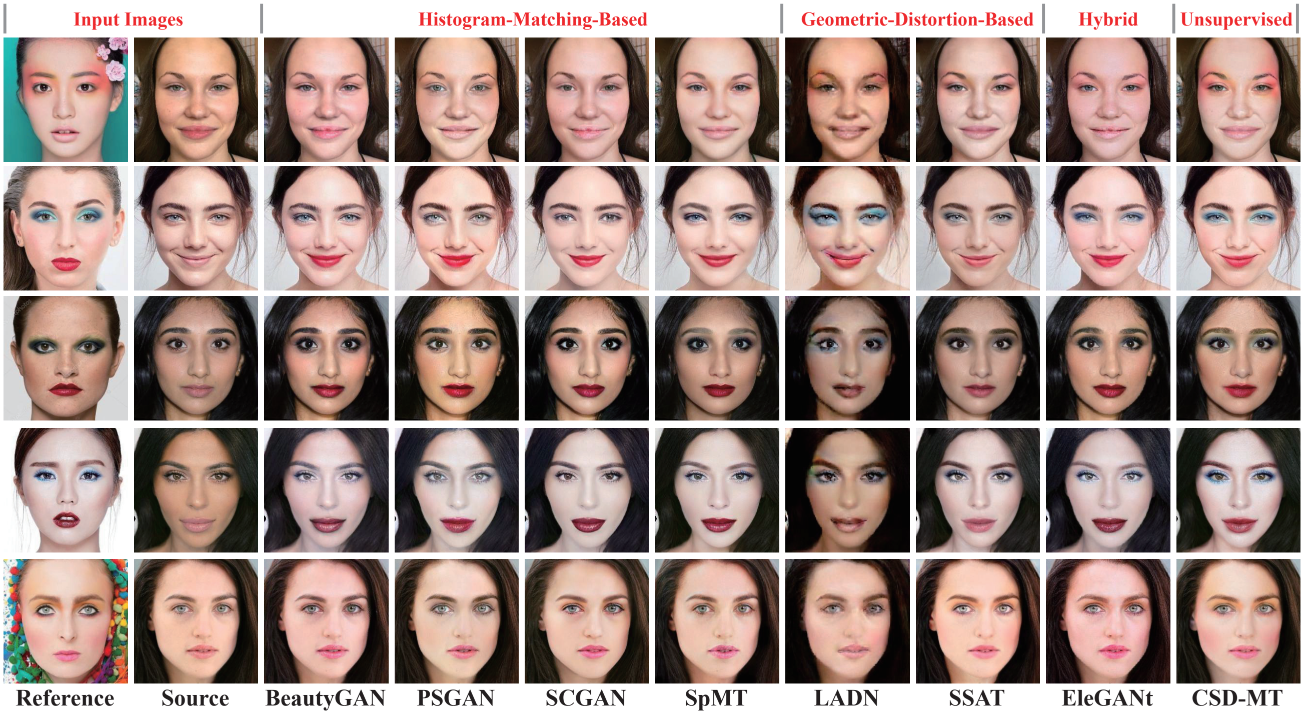


Figure 12. More qualitative comparisons between CSD-MT and state-of-the-art methods **under complex makeup styles**.

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. 4

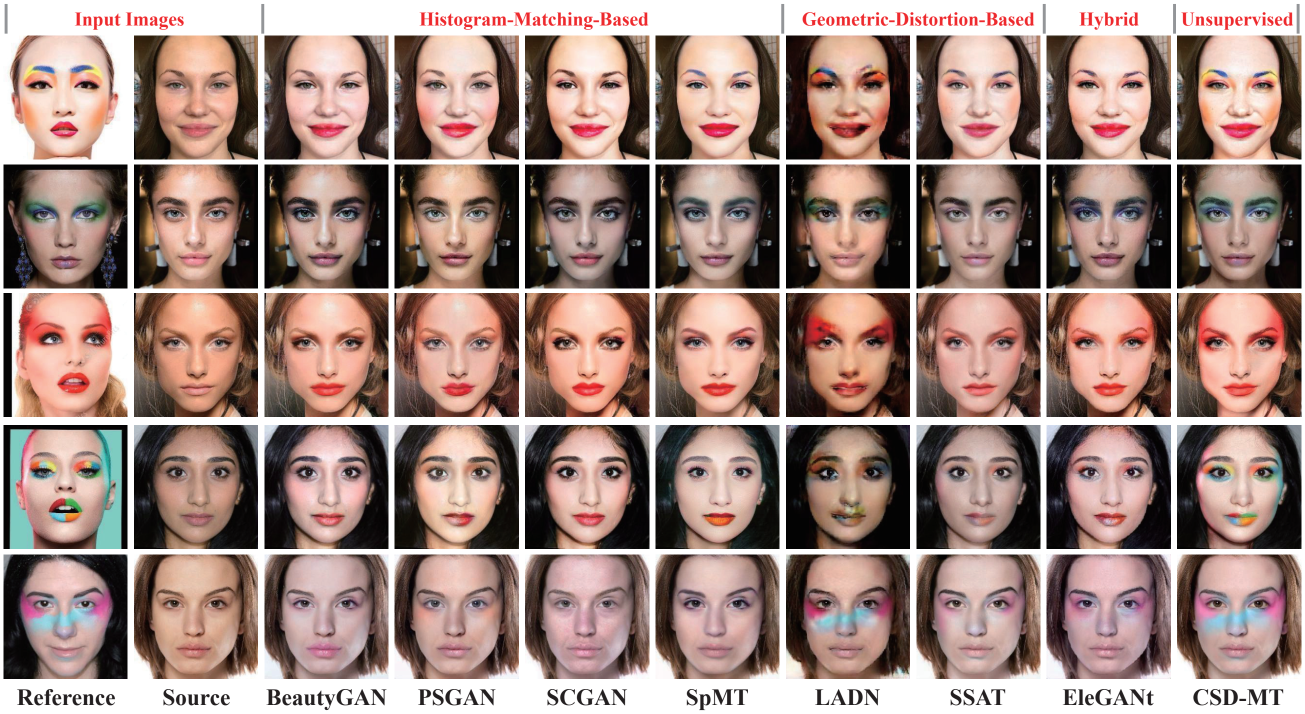


Figure 13. More qualitative comparisons between CSD-MT and state-of-the-art methods **under extreme makeup styles**.

[2] 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. 1

[3] 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. 1, 3

[4] 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. 1, 4

[5] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 3

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[7] 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. 1, 4

[8] 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*, pages 2325–2334, 2022. 1, 3, 4

[9] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1

[10] 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. 1, 4

[11] 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. 1

[12] Honglun Zhang, Wenqing Chen, Hao He, and Yaohui Jin. Disentangled makeup transfer with generative adversarial network. *arXiv preprint arXiv:1907.01144*, 2019. 3

[13] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023. 3

[14] Mingrui Zhu, Yun Yi, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. Semi-parametric makeup transfer via semantic-aware correspondence. *arXiv preprint arXiv:2203.02286*, 2022. 1, 4

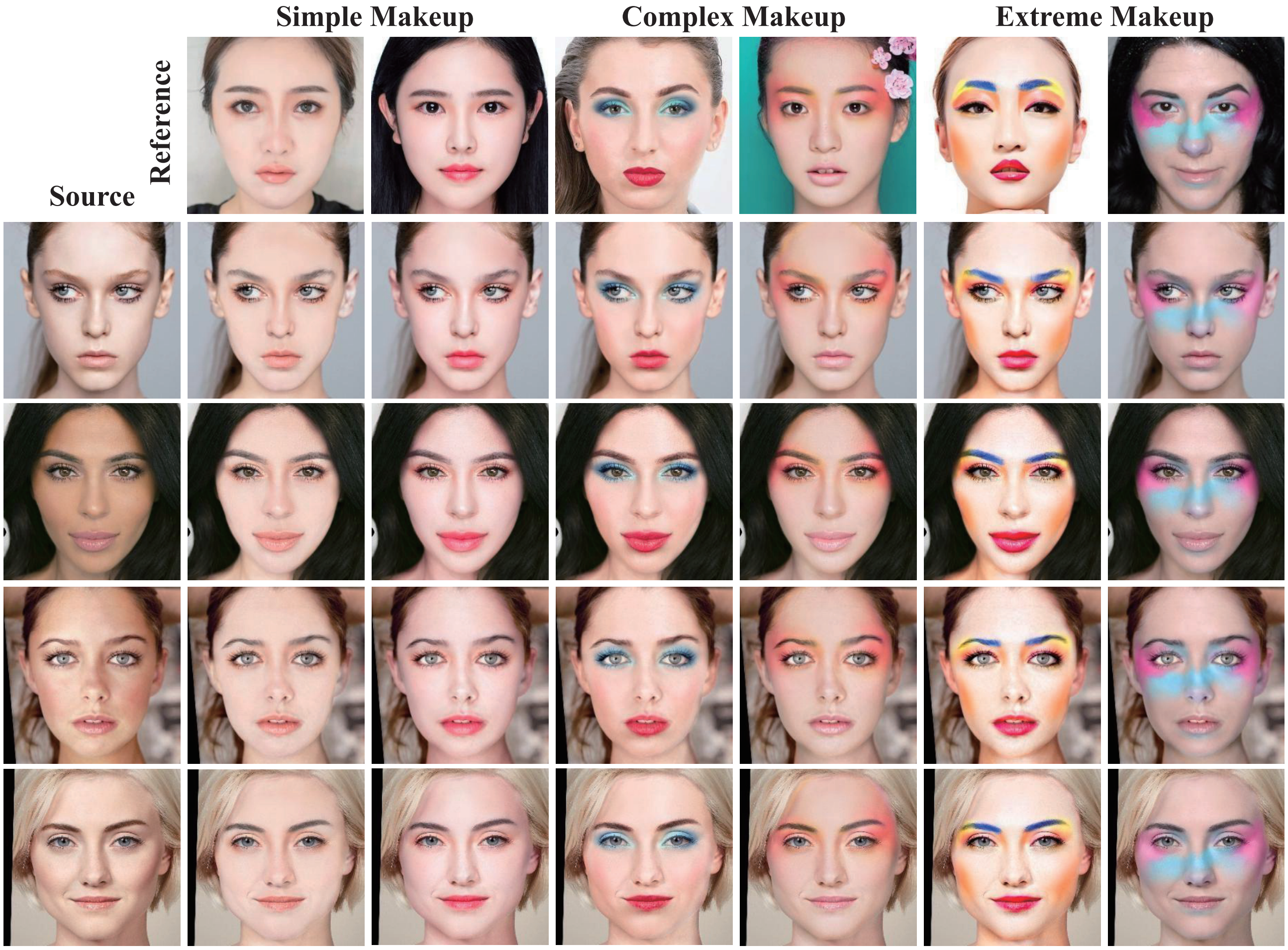


Figure 14. The makeup transfer results 1 of our CSD-MT under simple, complex, and extreme makeup styles.

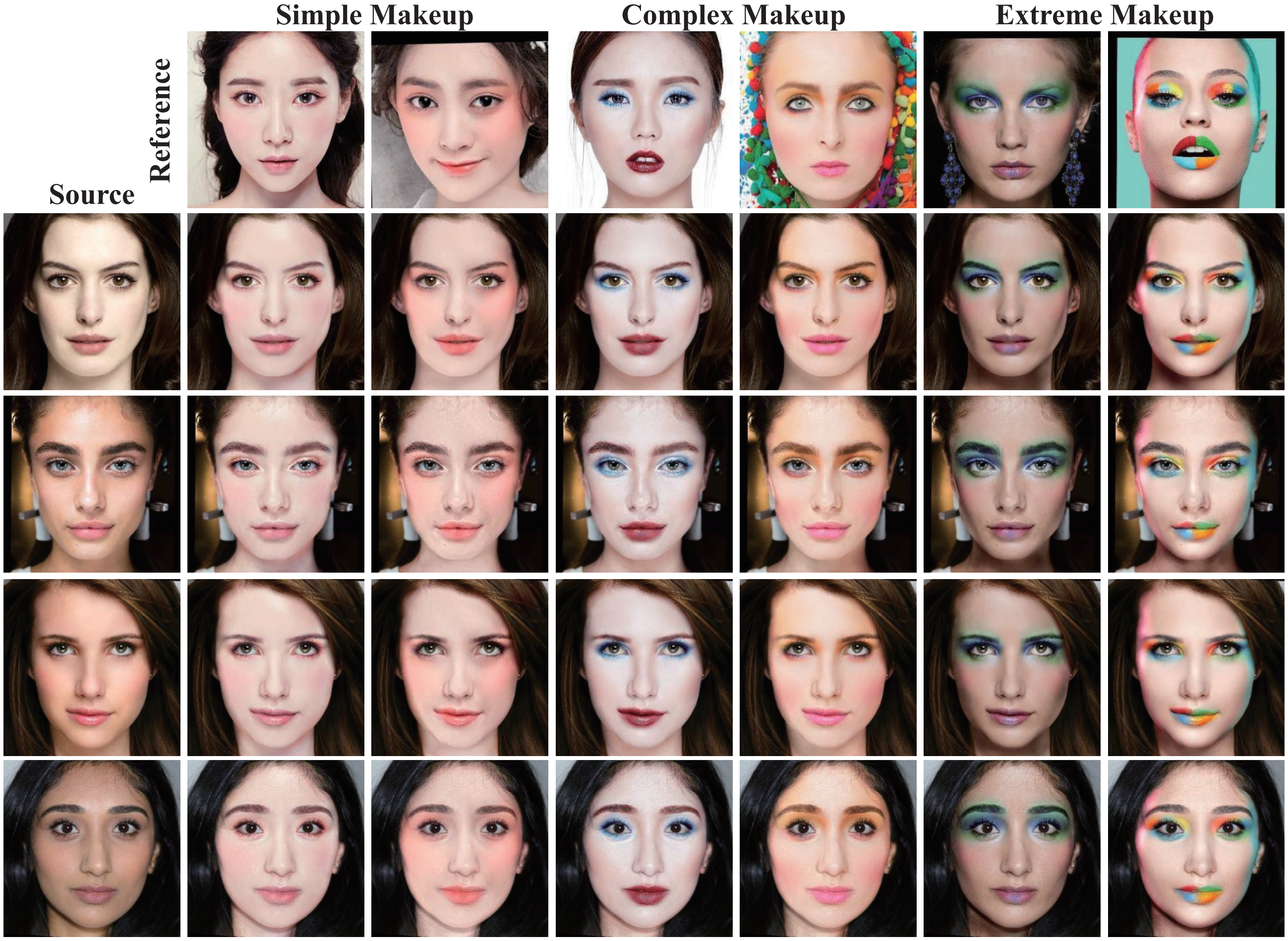


Figure 15. The makeup transfer results 2 of our CSD-MT under simple, complex, and extreme makeup styles.

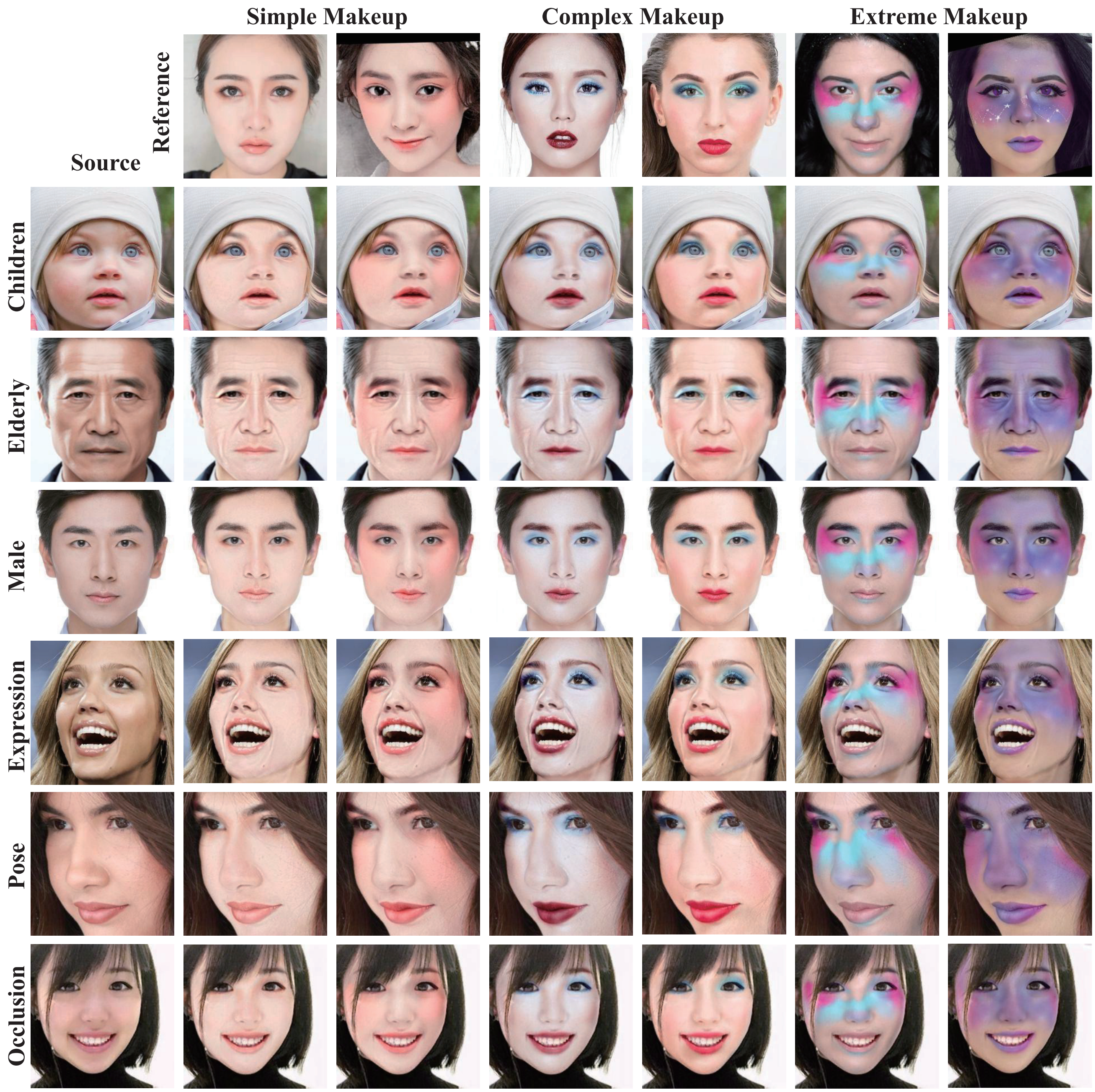


Figure 16. The robustness of CSD-MT in various complex scenarios.

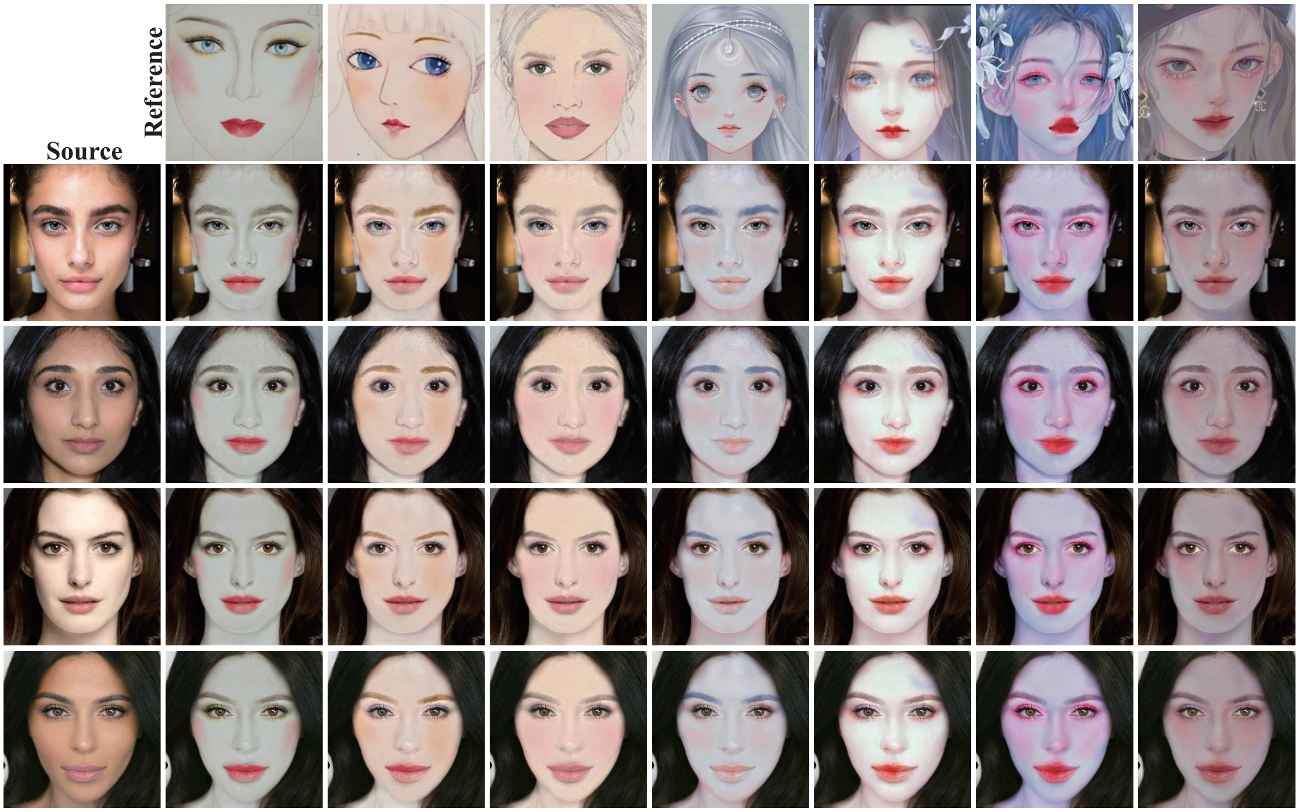


Figure 17. The generalization of CSD-MT in unsee anime makeup styles.

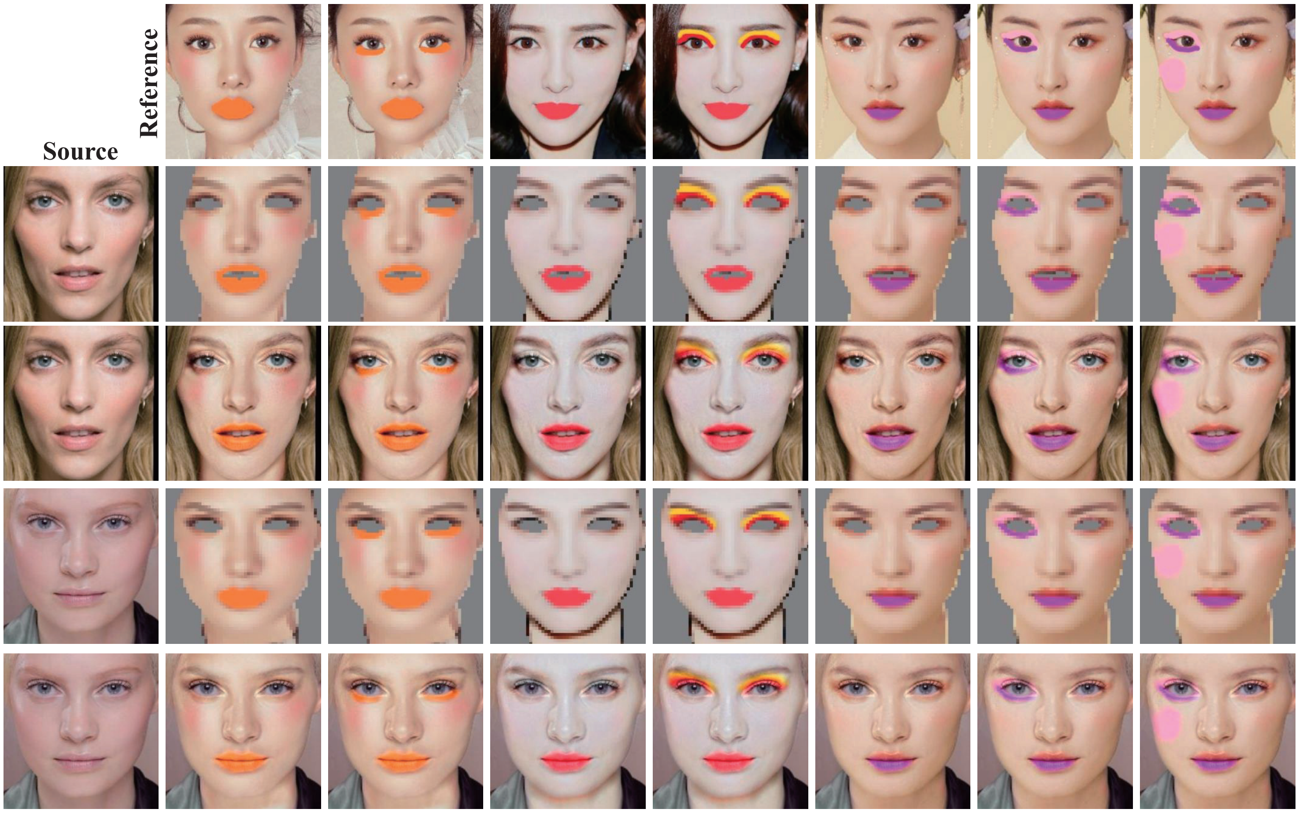


Figure 18. The controllability of CSD-MT in makeup editing. The deformed LF components are showcased to explain the makeup control mechanism of our approach.

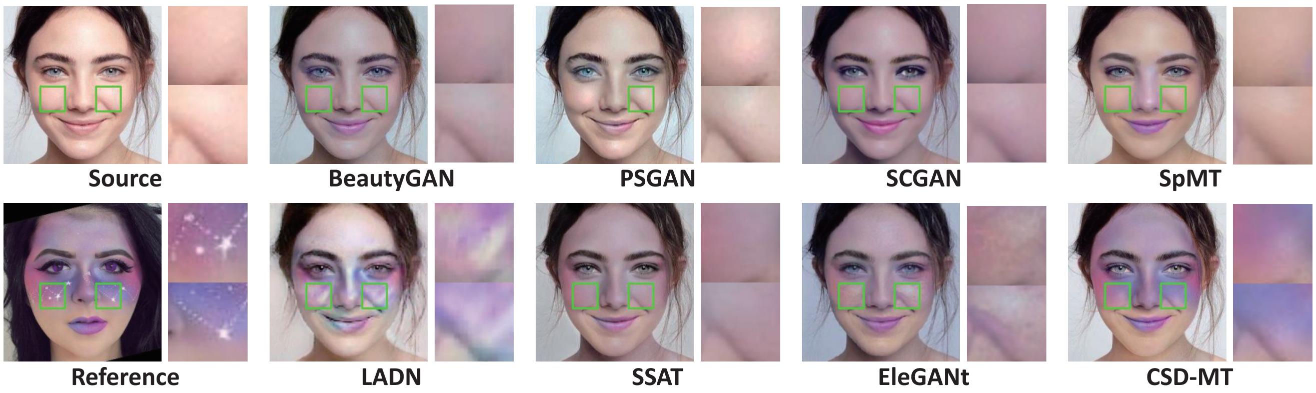


Figure 19. The limitation of our CSD-MT. We assume that the high-frequency (HF) component is more closely associated with the content details of face images. As a result, our CSD-MT is ineffective in accurately reproducing the boundaries of some extreme makeup styles.