SSAT: A Symmetric Semantic-Aware Transformer Network for Makeup Transfer and Removal

Zhaoyang Sun[1], **Yaxiong Chen**[1, 2], **Shengwu Xiong**[1, 3, *]

[1] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, 430070
[2] Wuhan University of Technology Chongqing Research Institute, Chongqing, 401122
[3] Sanya Science and Education Innovation Park of Wuhan University of Technology, Sanya, 572000
zhaoyangsun0304@outlook.com, chenyaxiong@whut.edu.cn, xiongsw@whut.edu.cn

Abstract

Makeup transfer is not only to extract the makeup style of the reference image, but also to render the makeup style to the semantic corresponding position of the target image. However, most existing methods focus on the former and ignore the latter, resulting in a failure to achieve desired results. To solve the above problems, we propose a unified Symmetric Semantic-Aware Transformer (SSAT) network, which incorporates semantic correspondence learning to realize makeup transfer and removal simultaneously. In SSAT, a novel Symmetric Semantic Corresponding Feature Transfer (SSCFT) module and a weakly supervised semantic loss are proposed to model and facilitate the establishment of accurate semantic correspondence. In the generation process, the extracted makeup features are spatially distorted by SSCFT to achieve semantic alignment with the target image, then the distorted makeup features are combined with unmodified makeup irrelevant features to produce the final result. Experiments show that our method obtains more visually accurate makeup transfer results, and user study in comparison with other state-of-the-art makeup transfer methods reflects the superiority of our method. Besides, we verify the robustness of the proposed method in the difference of expression and pose, object occlusion scenes, and extend it to video makeup transfer.

Introduction

Makeup transfer aims to transfer the makeup style of any reference image to the target image while preserving the identity of the target person. With the booming development of the cosmetics market, makeup transfer is widely demanded in many popular beautifying applications and has been received extensive attention in the computer vision and graphics. In the early years, traditional approaches (Guo and Sim 2009; Tong et al. 2007; Li, Zhou, and Lin 2015) mostly used image gradient editing or physical-based modification to realize makeup transfer. Recently, combining generative adversarial network (Goodfellow et al. 2014) and disentangled representation (Lee et al. 2018; Huang et al. 2018), many makeup transfer approaches (Li et al. 2018; Chang et al. 2018; Chen et al. 2019; Gu et al. 2019; Huang et al. 2020; Deng et al. 2021; Nguyen, Tran, and Hoai 2021) have made significant progress in extracting makeup style and generating realistic makeup results.

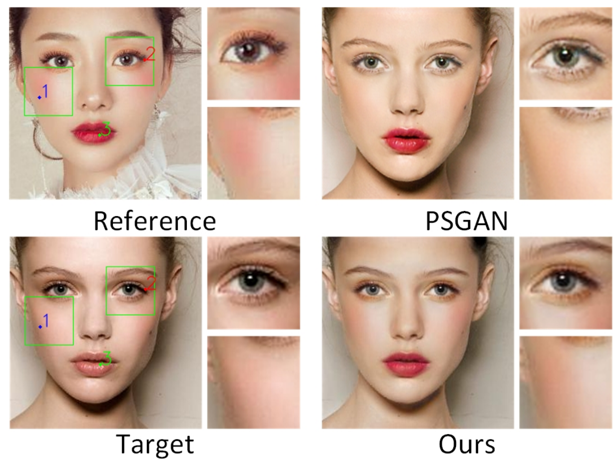

Figure 1: The diagram of our motivation and PSGAN comparison. The spatial mapping of the same color points between the target image and the reference image is the semantic correspondence expected to be established in this paper.

However, little attention has been paid to semantic correspondence, which plays an important role in makeup transfer. Consider the actual makeup process following the tutorial: 1) choose cosmetics of the same color (extract makeup style); 2) apply this cosmetics to the semantic corresponding position of the face (semantic correspondence), see Figure 1. Ignoring the semantic correspondence would result in the makeup style being transferred to the wrong position, such as lipstick leaking lips. At the same time, inaccurate semantic correspondence will lead to the averaging of makeup colors (Jiang et al. 2020), so that the resulting makeup style is visually different from the reference makeup. In addition, the robustness of expression and pose and the robustness of object occlusion are greatly reduced.

The motivation of this paper is to establish accurate semantic correspondence between the target and the reference facial images to improve the quality of makeup transfer. Note that the dense semantic correspondence of the makeup regions is established, not the sparse semantic correspon-

*Corresponding Author

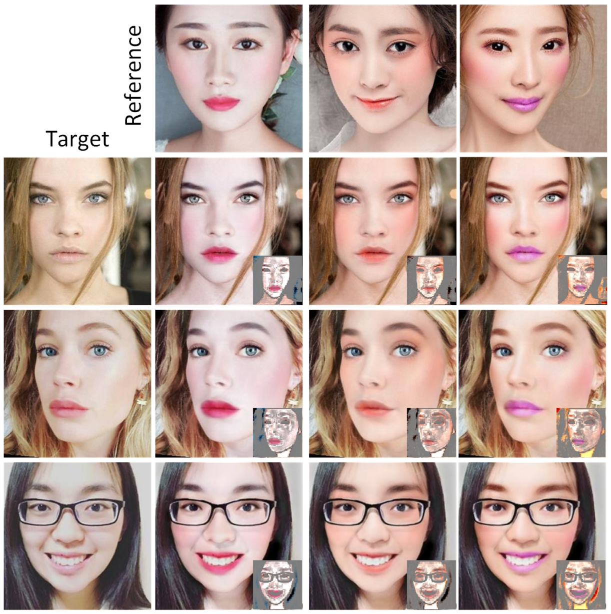

Figure 2: Our SSAT makeup transfer results. The diversity of makeup style, the difference of facial expression and pose, object occlusion (glasses or hair) are considered in the selected images. The semantic correspondence is in the lower right corner of the generated result, and the value of each spatial position of it is obtained by the different weighted sum of the reference image.

dence shown in the Figure1. For this purpose, a Symmetric Semantic Corresponding Feature Transfer (SSCFT) module is proposed to models the spatial position mapping between different images. Next there is no effective supervision for semantic correspondence, so the face parsing is introduced and a weakly supervised semantic loss is proposed. The experiments show that the semantic correspondence established by our method is adaptive, which ignores the semantic correspondence of makeup irrelevant regions (hair, background, glasses, inside eyes, inside mouth) and focuses on the makeup regions, see Figure 2.

To sum up, we propose a Symmetric Semantic-Aware Transformer network (SSAT) whose framework and process are shown in Figure 3. Compared with PSGAN (Jiang et al. 2020), there are several obvious differences: 1) The proposed method introduces the face parsing, instead of the feature points which sometimes has large error in side face. 2) Learnable network features instead of hand-designed features to establish semantic correspondence. 3) Using the symmetry of semantic correspondence, our method generates makeup transfer and removal results simultaneously, while PSGAN can only realize makeup transfer. The main contributions of this paper are summarized as follows:

- We propose a novel Symmetric Semantic-Aware Transformer (SSAT) network for makeup transfer and removal. Experiments verify the effectiveness of the transferring strategy in the real-world environment and generated results are of higher quality than state-of-the-art methods both qualitatively and quantitatively.
- A novel SSCFT module and a semantic loss are proposed to establishing accurate semantic correspondence, which significantly improves the quality of makeup transfer.
- Combined with face parsing, our method is more flexible and could realize partial makeup transfer. Meanwhile, we verify the robustness of the proposed method in the difference of expression and pose, object occlusion (glasses or hair) scenes and extend it to video makeup transfer.

Related Work

Makeup Transfer

In recent years, makeup transfer has been extensively studied. BeautyGAN (Li et al. 2018) addressed the makeup transfer and removal task by incorporating both global domain adaptation loss and local instance-level makeup loss in an dual input/output GAN. PairedCycleGAN (Chang et al. 2018) extended the CycleGAN (Zhu et al. 2017) to asymmetric networks to enable transferring specific makeup style. LADN (Gu et al. 2019) and CPM (Nguyen, Tran, and Hoai 2021) focused on the complex/dramatic makeup styles transfer. Introducing facial feature points, PSGAN (Jiang et al. 2020; Liu et al. 2021) proposed a pose and expression robust spatial-aware GAN for makeup transfer. Recently, SOGAN (Lyu et al. 2021) explored the shadow and occlusion robust and (Wan et al. 2021) realized makeup transfer from the perspective of face attribute editing. Inspired by StyleGAN (Karras, Laine, and Aila 2018), SCGAN (Deng et al. 2021) proposed a style-based controllable GAN model. Unlike the above methods, the core idea of this paper is to establish accurate semantic correspondence to improve the quality of makeup transfer.

Semantic Correspondence

In recent years, CNN-based features (Simonyan and Zisserman 2015; Krizhevsky, Sutskever, and Hinton 2017) have been proved to be a powerful tool to express high-level semantics. Recently, (Liao et al. 2017; Zhang et al. 2020) proposed a technique for visual attribute transfer across images using semantic correspondence. The exemplar-based colorization methods (He et al. 2018; Lee et al. 2020) calculated the semantic correspondence between the target image and the exemplar image, then transferred the color with the closest semantic similarity to the target image. Inspired by the recent exemplar-based image colorization (He et al. 2018; Lee et al. 2020), our work is expected to transfer the makeup style with the closest semantic similarity to the target image.

Our Approach: SSAT

Formulation

Our goal is to transfer the makeup style from an arbitrary makeup reference image to a non-makeup target image. Here, $X \subset \mathcal{R}^{H \times W \times 3}$ refers to non-makeup image domain, $Y \subset \mathcal{R}^{H \times W \times 3}$ refers to makeup image domain. Given a target image $x_t \in X$ and a reference image $y_r \in Y$, the goal of makeup transfer is learning a mapping function:

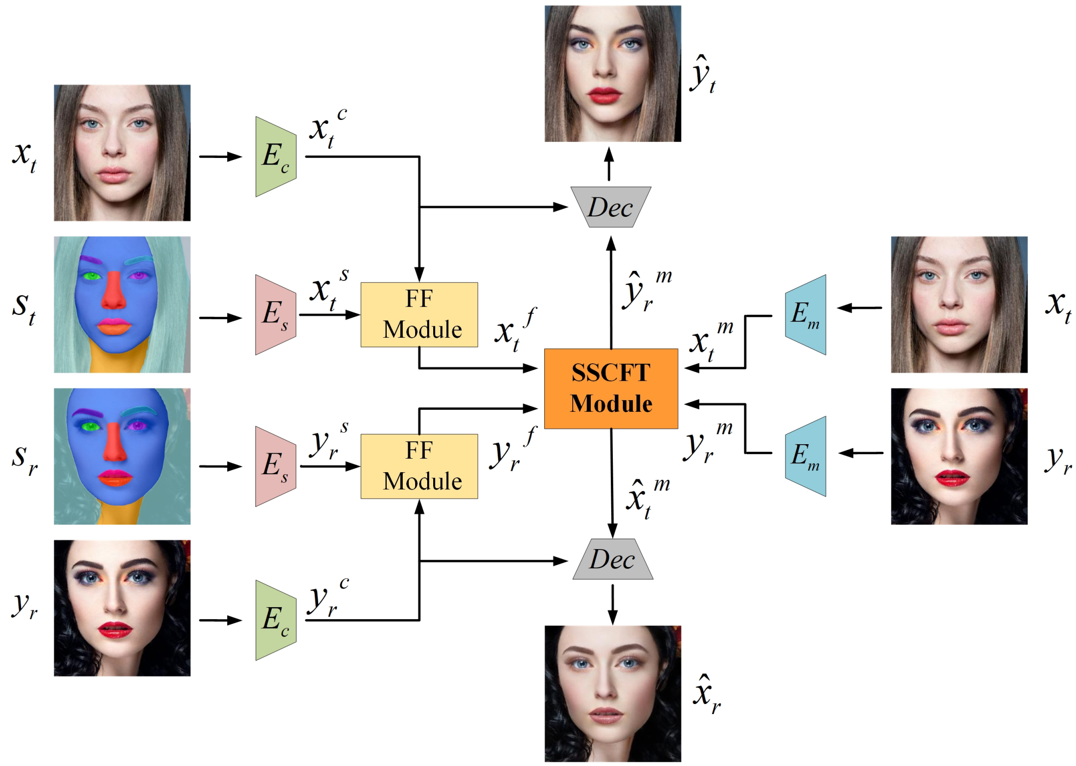

Figure 3: The proposed Symmetric Semantic-Aware Transformer (SSAT) network. The process goes through the following steps: 1) Content encoder E_c and makeup encoder E_m decompose target image x_t and reference image y_r respectively, $x_t^c = E_c(x_t), x_t^m = E_m(x_t), y_r^c = E_c(y_r), y_r^m = E_m(y_r)$. Meanwhile, face parsing s_t, s_r is introduced and semantic features are extracted by using semantic encoders E_s: $x_t^s = E_s(s_t), y_r^s = E_s(s_r)$. 2) The Feature Fusion (FF) module fuses content features and semantic features to obtain richer features for semantic correspondence, $x_t^f = FF(x_t^c, x_t^s), y_r^f = FF(y_r^c, y_r^s)$. 3) The Symmetric Semantic Corresponding Feature Transfer (SSCFT) module distorts makeup features spatially according to the semantic correspondence established by x_t^f and y_r^f, and outputs $\hat{y}_r^m, \hat{x}_t^m = SSCFT(x_t^f, y_r^f, x_t^m, y_r^m)$. 4) Distorted makeup features $\hat{y}_r^m$ of the reference image are embedded in the content features x_t^c of the target image to generate makeup transfer result $\hat{y}_t = Dec(x_t^c, \hat{y}_r^m)$. Similarly, the makeup removal result $\hat{x}_r = Dec(y_r^c, \hat{x}_t^m)$.

$\Phi : x_t, y_r \to \hat{y}_t$, where $\hat{y}_t \in Y$ has the makeup style with y_r while preserving the identity of x_t. For makeup removal, it is assumed that the non-makeup image is a special case of the makeup image (Sun et al. 2020), which unifies makeup transfer and makeup removal. Therefore, the goal of makeup removal is learning a mapping function: $\Phi : y_r, x_t \to \hat{x}_r$, where $\hat{x}_r \in X$ has the makeup style with x_t while preserving the identity of y_r. In this paper, the only difference between makeup transfer and removal is whether the reference image is a non-makeup image or a makeup image.

SSAT

The overall framework of Symmetric Semantic-Aware Transformer network (SSAT) is shown in Figure 3, which consists of three encoders, a FF module, a SSCFT module and a decoder Dec. Next, the function of each module will be introduced in detail.

Encoders The one main problem of makeup transfer stems from the difficulty of extracting the makeup latent features, which are required to be disentangled from other makeup irrelevant features. Here, this problem is referred to as content-style separation (Huang et al. 2018; Lee et al. 2018). So a content encoder E_c and a makeup encoder E_m are designed to extract content features and makeup features respectively:

$$x_t^c = E_c(x_t), x_t^m = E_m(x_t) \tag{1}$$

$$y_r^c = E_c(y_r), y_r^m = E_m(y_r) \tag{2}$$

Experiments have found that it is difficult to establish accurate semantic correspondences only with content features. Therefore, face parsing (Yu et al. 2018) is introduced and semantic features are extracted by using semantic encoders E_s:

$$x_t^s = E_s(s_t), y_r^s = E_s(s_r) \tag{3}$$

where $s_t \in \mathcal{R}^{H \times W \times L}$ and $s_r \in \mathcal{R}^{H \times W \times L}$ refer to binary face parsing of the target image and the reference image, respectively. L is the number of different semantic regions, which is set 18 in our experiments. Next, content features and semantic features will cooperate to establish semantic correspondence.

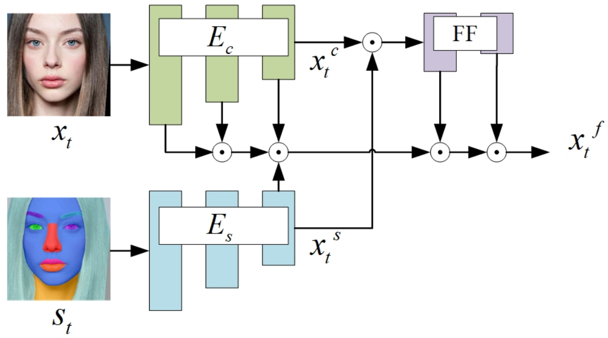

Figure 4: The illustration of feature fusion (FF) module. FF fuses content features and semantic features to obtain richer features for semantic correspondence. $\odot$ refers to the channel-wise concatenation operator.

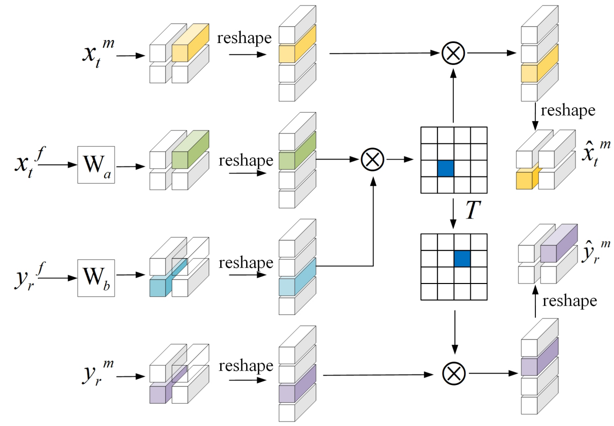

Figure 5: The illustration of symmetric semantic corresponding feature transfer (SSCFT) module. W_a, W_b refer to the linear transformation matrix into x_t^f and y_r^f, respectively. $\otimes$ refers to the matrix multiplication operator. T refers to the matrix transpose operator. After this module, the output distorted makeup features $\hat{x}_t^m, \hat{y}_r^m$ are semantic aligned with y_r^c, x_t^c in semantic.

FF In the FF module, content features and semantic features are fused to obtain richer features for feature matching, see Figure 4. Take the target image as an example, after obtaining content features and semantic features, they are connected along the channel dimensions and fed into the FF module. FF consists of N convolutional layers and outputs N high-level features $(f^1, f^2, \cdots, f^N)$. Besides, the low-level features extracted by E_c and E_s are also combined. E_c and E_s consist of M convolutional layers, producing M low-level features $(C^1, C^2, \cdots, C^M)$ and $(S^1, S^2, \cdots, S^M)$, where $C^M = x_t^c, S^M = x_t^s$. Each layer of content features is selected, but the last layer of semantic features is selected. Now, we downsample each of C^m and upsample each of f^n to match the spatial size S^M, forming the final fusion features x_t^f:

$$x_t^f = [\phi(C^1); \phi(C^2); \cdots; C^M; S^M; \\ \varphi(f^1); \varphi(f^2); \cdots; \varphi(f^N)] \tag{4}$$

where ϕ denotes a spatially downsampling function of an input C^m of different size to the size of C^M or S^M. Similarly, φ denotes a spatially upsampling function. ";" denotes the channel-wise concatenation operator. In this manner, the output fusion feature x_t^f combines the low-level and high-level features, while ensuring the accuracy of semantic correspondence. With the same operation, we obtain the fusion features y_r^f of the reference image.

SSCFT The SSCFT module is inspired by colorization (He et al. 2018) and combines the symmetry of semantic correspondence. In makeup transfer task, the semantic correspondence is a one-to-one mapping relationship. That is, point A corresponds to point B. In turn, point B also corresponds to point A. And the symmetry of this semantic relationship can be applied cleverly to the anti-task of makeup transfer, makeup removal. See Figure 5 for the framework of the proposed SSCFT module.

Firstly, we reshape x_t^f into $\hat{x}_t^f = [x_1, x_2, \cdots, x_{hw}] \in \mathcal{R}^{d \times hw}$, where $x_i \in \mathcal{R}^d$ indicates feature variables of the i^{th} location of $\hat{x}_t^f$. Then we obtain channel-wise centralized features $\hat{x}_i$ and $\hat{y}_j$ to make the learning more stable (He et al. 2018), where $\hat{x}_i = x_i - mean(x_i), \hat{y}_j = x_y - mean(y_j)$. Given $\hat{x}_i$ and $\hat{y}_j$, SSCFT computes a semantic correlation matrix $\mathcal{A} \in \mathcal{R}^{hw \times hw}$, whose element $a_{i,j}$ is computed by the scaled dot product:

$$a_{i,j} = \frac{\hat{x}_i^T \cdot \hat{y}_j}{\|\hat{x}_i\| \|\hat{y}_j\|} \tag{5}$$

After that, we distort the reference makeup features y_r^m towards x_t^c according to the correlation matrix $\mathcal{A}$. The weighted sum of $\hat{y}_r^m$ is calculated to approximate the makeup sampling from y_r^m :

$$\hat{y}_r^m(i) = \sum_j softmax_j(a(i,j) \cdot \sigma) \cdot y_r^m(j) \tag{6}$$

where σ controls the sharpness of the softmax and we set its default value as 100. Now the distorted makeup features $\hat{y}_r^m$ of reference image are aligned with the content features x_t^c of target image in semantic. In the same way, we obtain the distorted makeup features $\hat{x}_t^m$, which aligns with the content features y_r^c. Note that this step makes our method robust to expression, pose, object occlusion and produce more accurate makeup transfer results.

Dec Finally, we employ the spatially-adaptive denormalization (SPADE) block (Park et al. 2019) to project the distorted makeup styles $\hat{y}_r^m, \hat{x}_t^m$ to content features x_t^c, y_r^c for makeup transfer and removal.

$$\hat{y}_t = Dec(x_t^c, \hat{y}_r^m) \tag{7}$$

$$\hat{x}_r = Dec(y_r^c, \hat{x}_t^m) \tag{8}$$

where $\hat{y}_t$ is the makeup transfer result and $\hat{x}_r$ is the makeup removal result.

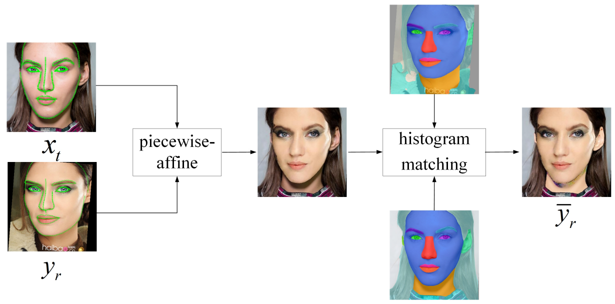

Figure 6: The process of generating pseudo-paired data.

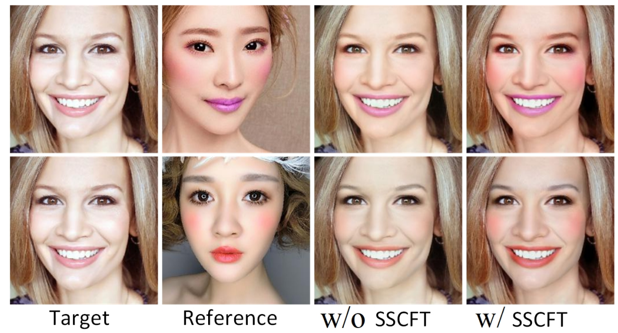

Figure 7: Ablation study of SSCFT module.

Objective

In total, there are four loss functions used for network SSAT end-to-end training. The overall loss is as follows:

$$\begin{aligned} L_{overall} = &\lambda_{sem} L_{sem} + \lambda_{makeup} L_{makeup} + \\ &\lambda_{rec} L_{rec} + \lambda_{adv} L_{adv} \end{aligned} \tag{9}$$

Semantic Loss A weakly supervise semantic loss is proposed to establish semantic correspondence. The idea is that semantic correspondence should only exist between semantic regions of the same class:

$$L_{sem} = \|s_t - \hat{s}_r\|_1 + \|s_r - \hat{s}_t\|_1 \tag{10}$$

where $s_t \in \mathcal{R}^{H\times W\times L}$ and $s_r \in \mathcal{R}^{H\times W\times L}$ refer to binary face parsing, which only have the values 0 and 1. L is the number of semantic classes, $\hat{s}_r(i) = \sum_j softmax_j(a(i,j)\cdot \sigma)\cdot s_r(j)$, $\hat{s}_t(i) = \sum_j softmax_j(a(i,j)\cdot\sigma)\cdot s_t(j)$, and $\|\cdot\|_1$ refers to L_1 loss. For dense semantic correspondence, the L_{sem} is only a crude region constraint, but experiments show that it plays an important role.

Makeup Loss Inspired by (Chang et al. 2018), we generate pseudo pairs of data $\bar{x}_r$ and $\bar{y}_t$ according to face feature points to train the network, see Figure 12. The process is described in detail in the supplementary materials. Here, we introduce the SPL loss (Sarfraz et al. 2019) instead of the L_1 loss to guide the makeup transfer and removal:

$$L_{makeup} = SPL(\hat{y}_t, x_t, \bar{y}_t) + SPL(\hat{x}_r, y_r, \bar{x}_r) \tag{11}$$

Take the makeup transfer as an example, SPL constrains the gradient consistency between $\hat{y}_t$ and x_t to ensure the identity of the target image, and restricts the color consistency between $\hat{y}_t$ and $\bar{y}_t$ to guide the makeup transfer. Note that L_{sem} constrains the corresponding region and does not penalize one-to-many mappings of the same class. While L_{makeup} guides makeup transfer, it implicitly constrains the one-to-one mapping in the makeup regions.

Reconstruction Loss We feed the x_t^c and x_t^m into Dec to generate x_t^{self}, and y_r^c and y_r^m into Dec to generate y_r^{self}, which should be identical to x_t and y_r. Here, we introduce Cycle loss (Zhu et al. 2017) to ensure that the image does not lose information during the decoupling process of makeup features and content features. So we feed the makeup removal result $\hat{x}_r$ and makeup transfer $\hat{y}_t$ result into SSAT again to obtain the x_t^{cycle} and y_r^{cycle}, which should also be identical to x_t and y_r. We use L1 loss to encourage such reconstruction consistency:

$$\begin{aligned} L_{rec} = &\|x_t - x_t^{self}\|_1 + \|y_r - y_r^{self}\|_1 + \\ &\|x_t - x_t^{cycle}\|_1 + \|y_r - y_r^{cycle}\|_1 \end{aligned} \tag{12}$$

Adversarial Loss Two discriminators D_X and D_Y are introduced for the non-makeup domain and makeup domain, which try to discriminate between real samples and generated images and help the generator synthesize realistic outputs. The least square loss (Mao et al. 2017) is used for steady training:

$$\begin{aligned} L_{adv} = &\mathbf{E}_{x_t}[(D_X(x_t))^2] + \mathbf{E}_{\hat{x}_r}[(1-D_X(\hat{x}_r))^2] + \\ &\mathbf{E}_{y_r}[(D_Y(y_r))^2] + \mathbf{E}_{\hat{y}_t}[(1-D_Y(\hat{y}_t))^2] \end{aligned} \tag{13}$$

Experiment

Dataset and Implementation Details

For the dataset, we randomly selected 300 non-makeup images as target images and 300 makeup images as reference images from the Makeup Transfer dataset (Li et al. 2018). Using the proposed generation method, a total of 180,000 pairs of pseudo-paired data are generated for makeup transfer and removal. During the training, all trainable parameters are initialized normally, and the Adam optimizer with $\beta_1 = 0.5, \beta_2 = 0.999$ is employed for training. We set $\lambda_{sem} = 1, \lambda_{makeup} = 1, \lambda_{rec} = 1, \lambda_{adv} = 1$ for balanceing the different loss functions. The SSAT is implemented by MindSpore [1]. And the model is trained for 300,000 iterations on one single Nvidia 2080Ti GPU. The learning rate is fixed as 0.0002 during the first 150,000 iterations and linearly decays to 0 over the next 150,000 iterations. The batchsize is set 1. See the supplementary materials for the specific network structure parameters. Code will be available at https://gitee.com/sunzhaoyang0304/ssat-msp.

Ablation Studies

Effects of SSCFT The motivation of this paper is to establish semantic correspondence to improve the quality of makeup transfer. In order to verify our idea, we remove the SSCFT module to analyze the effects of the role of semantic

[1] Mindspore. https://www.mindspore.cn/

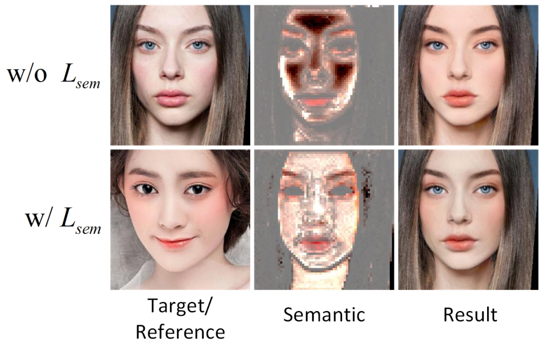

Figure 8: Ablation study of L_{sem}.

correspondence in makeup transfer. In this case, we directly skip the SSCFT module and input the features into the decoder. The comparison results are shown in the Figure 7. Without SSCFT, the resulting makeup style is significantly lighter than the reference makeup. With the addition of SSCFT, the result is more similar visually to the reference makeup , especially eye shadow and blush.

Effects of L_{sem} The lack of effective supervision hinders the establishment of semantic correspondence. To solve this problem, we introduce face parsing and propose a weakly supervised semantic loss L_{sem}. In order to verify their effect, we remove L_{sem} and the face parsing, use the remaining loss to train the network. The comparison of semantic and makeup transfer results is shown in the Figure 8. Compared without L_{sem}, the semantic result using L_{sem} is more accurate and the boundary between makeup area and non-makeup area is also clearer. For the makeup transfer result, the blush is mapped to the wrong semantic position, spreading over the entire face in the result without using L_{sem}.

Comparisons to Baselines

To verify the superiority of our makeup transfer strategy, we choose three state-of-the-art makeup transfer approaches, BeautyGAN (Li et al. 2018), PSGAN (Jiang et al. 2020), SCGAN (Deng et al. 2021), as our comparison benchmark. We skip some baselines such as LADN (Gu et al. 2019) and CPM (Nguyen, Tran, and Hoai 2021), because they focus on the complex/dramatic makeup styles transfer.

Qualitative Comparison Here, three scenes that often appear in real life are selected for comparison, frontal face, different expressions and poses, and object occlusion. The qualitative comparison has been shown in Figure 9. BeautyGAN produces a realistic result, but fails to transfer eye shadow and blush. Although PSGAN designs semantic correspondence modules, its accuracy is limited by hand-designed features. The imprecise semantic correspondence results in a lighter eye and blush makeup style, which is visually different from the reference makeup. Similar to BeautyGAN, the transfer of eye makeup and blush fails in SCGAN. Meanwhile, the results of SCGAN are too smooth, and slightly change makeup irrelevant information, such as the glasses in the sixth row. On the contrary, the makeup

Methods	Rank 1	Rank 2	Rank 3	Rank 4
BeautyGAN	3.0%	3.9%	34.5%	58.6%
PSGAN	11.3%	73.2%	11.3%	4.1%
SCGAN	3.6%	15.3%	48.0%	33.2%
SSAT	82.1%	7.6%	6.2%	4.1%

Table 1: User Study.

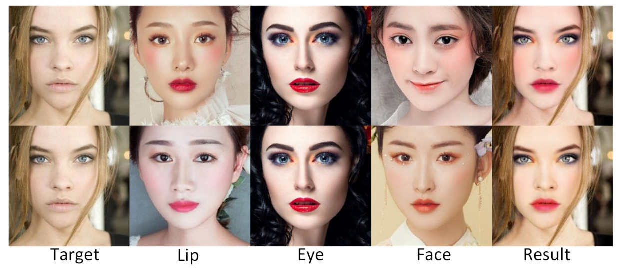

Figure 10: The partial makeup transfer results. The last column is the partial makeup transfer results, which receive personal identity from the first column, the lips style from the second column, the eyes style from the third column and the face style from the fourth column.

style of our results is highly similar to the reference makeup, whether it is lipstick, eye shadow, blush. The semantic results explain why our method has a better makeup transfer effect and why our method is robust to expression, pose and object occlusion.

Quantitative Comparison How quantitative evaluation makeup transfer is still a field that needs to explore. Here, we conduct user study to compare different methods quantitatively. We randomly generate 30 results of makeup transfer using four methods respectively. 45 volunteers are asked to rank the results based on the realism and the similarity of makeup styles. To be fair, the results in each selection are also randomly arranged. As shown in the Table 3, our SSAT outperforms other methods by a large margin. Our method achieves the highest selection rate of 82.1% in Rank 1.

Partial Makeup Transfer

Partial makeup transfer refers to transfer partial makeup of the reference image. The distorted makeup features extracted by our method is accurately distributed according to the spatial semantic position of the target image, making it possible to integrate the partial makeup from different reference images, as shown in the Figure 10.

$$\hat{y}_t^{part} = Dec(x_t^c, \sum_i (\hat{y}_r^{m_i} \cdot Mask_i)) \tag{14}$$

where $i \in \{Lip, Eye, Face\}$ in our experiment, $\hat{y}_r^{m_i}$ means the distorted makeup features extracted are from different reference images, $Mask_i$ represents a binary mask related to the makeup area.

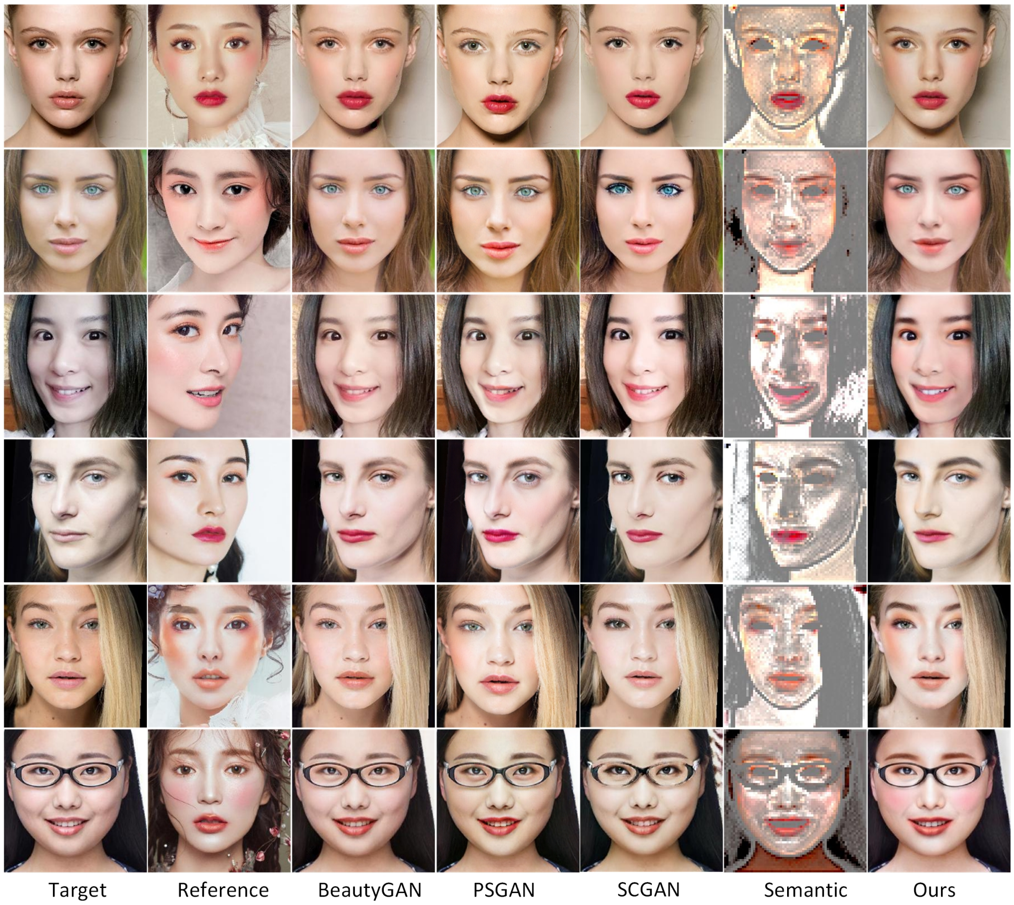

Figure 9: Comparison with state-of-the-art methods. The first two rows: frontal face, the middle two rows: different expressions and poses, the last two rows: object occlusion (glasses or hair).

Figure 11: The video makeup transfer results.

Video Makeup Transfer

Video makeup transfer is a very challenging task, which has high requirements for the quality of generated images and the accuracy of semantic correspondence. We download a video from the Internet and decompose it frame by frame, then apply the SSAT method, and finally integrate the resulting images into a video. We chose PSGAN as the comparison baseline, because other methods don't consider semantic correspondence. See Figure 11, the results produced by PSGAN are visually different from the reference makeup and cause flickering and discontinuity. In contrast, Our SSAT achieves smooth and accurate video makeup transfer results.

Conclusion

Different from other methods, we focus on semantic correspondence learning, propose the SSCFT module and a semantic loss, then integrate them into one Symmetric Semantic-Aware Transformer network (SSAT) for makeup transfer and removal. The experiment verified that semantic correspondence significantly improved the quality of makeup transfer visually as expected. The comparison with other methods demonstrates that our method achieves state-of-the-art makeup transfer results. In addition, benefits from precise semantic correspondence, our method is robust to the difference of expression and pose, object occlusion and can achieve partial makeup transfer. Moreover, we extend SSAT to the field of video makeup transfer, generating smooth and stable results. However, the computational complexity of proposed SSCFT is quadratic to image size, the focus of our later work is to solve this problem.

Acknowledgments

This work was in part supported by NSFC (Grant No. 62176194) and the Major project of IoV, Technological Innovation Projects in Hubei Province (Grant No.2020AAA001, 2019AAA024) and Sanya Science and Education Innovation Park of Wuhan University of Technology (Grant No. 2020KF0054), the National Natural Science Foundation of China under Grant 62101393 and the Open Project of Wuhan University of Technology Chongqing Research Institute(ZL2021–6).

Supplementary Materials

Generation of Pseudo-paired Data

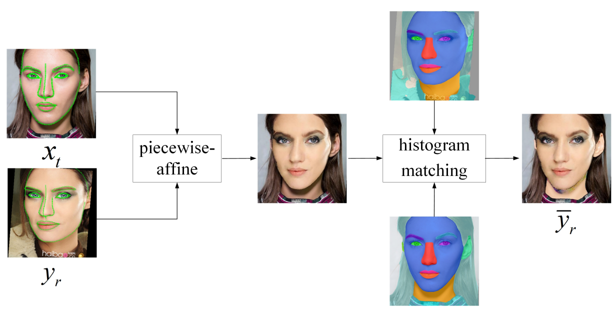

Figure 12: The process of generating pseudo-paired data.

Pseudo-paired data is used in the L_{makeup}, and the generation process is shown in Figure 12. Firstly, Face ++ API interface [2] is used to obtain the feature points of the input face image. Then, piecewise-affine transformation distorts the reference image into the target image according to the position of feature points. In this step, the inside of the eyes and the mouth remain unchanged, because these two areas need to be preserved during makeup transfer. Finally, histogram matching are applied to change the appearance of the ears and neck according to face parsing. Note that the warping results $\bar{y}_r$ sometimes possess artifacts, which can be fixed by SSAT in the generated results. For the dataset, we randomly selected 300 non-makeup images as target images and 300 makeup images as reference images from the Makeup Transfer dataset. Using the proposed generation method, a total of 180,000 pairs of pseudo-paired data are generated for makeup transfer and removal. During the training, images are resized to 286×286, randomly cropped to 256×256, horizontally flipped with probability of 0.5, and randomly rotated from -30 degrees to +30 degrees for data augmentation.

Network Structure

For the network structure, the encoders consist of stacked convolution blocks, which contain convolution, Instance normalization, and ReLU activation layers. In the decoder, the SPADE is applyed to embed the makeup features into the content features, with each SPADE followed by a residual block. For the discriminator, the network structure follows the multi-scale discriminator architecture. The specific network structure parameters will be given in detail in the table below. I, O, K, P, and S denote the number of input channels, the number of output channels, a kernel size, a padding size, and a stride size, respectively. The network architecture of Encoder E_c and E_m has been shown in table 2. The network architecture of Encoder E_s has been shown in table 3 The network architecture of FF module has been shown in table 4 The network architecture of Encoder Dec has been shown in table 5

[2]https://www.faceplusplus.com.cn/dense-facial-landmarks/

Layer	E_c and E_m
L1	Conv(I:3,O:64,K:7,P:3,S:1), Instance Normalization, Leaky ReLU:0.2
L2	Conv(I:64,O:128,K:3,P:1,S:2), Instance Normalization, Leaky ReLU:0.2
L3	Conv(I:128,O:256,K:3,P:1,S:2), Instance Normalization, Leaky ReLU:0.2

Table 2: The network architecture of Encoder E_c and E_m.

Layer	E_s
L1	Conv(I:18,O:32,K:7,P:3,S:1), Instance Normalization, Leaky ReLU:0.2
L2	Conv(I:32,O:64,K:3,P:1,S:2), Instance Normalization, Leaky ReLU:0.2
L3	Conv(I:64,O:128,K:3,P:1,S:2), Instance Normalization, Leaky ReLU:0.2

Table 3: The network architecture of Encoder E_s.

Makeup Removal

The makeup image as the target images and the non-makeup images as the reference images, then feed them into SSAT, the makeup removal results can be obtained, see Figure 13. In addition, more of our makeup transfer results are shown in Figure 15.

Makeup Style Interpolation

Adjusting the shade of makeup style is an essential function of existing makeup applications. Due to the separation of makeup features, our method could generate continuous makeup transfer results by interpolating makeup features, see Figure 14 . The formula is described as follows:

$$\hat{y}_t^{adjust} = Dec(x_t^c, (\hat{y}_r^{m_1} \cdot \alpha_1 + \hat{y}_r^{m_2} \cdot \alpha_2)) \qquad (15)$$

where $\alpha_1 + \alpha_2 = 1$. The closer α_1 is to 1, the closer the resulting makeup style is to $\hat{y}_r^{m_1}$. The closer α_2 is to 1, the closer the resulting makeup style is to $\hat{y}_r^{m_2}$.

Layer	FF
L1	Conv(I:384,O:512,K:3,P:1,S:2), Instance Normalization, Leaky ReLU:0.2
L2	Conv(I:512,O:512,K:3,P:1,S:1), Instance Normalization, Leaky ReLU:0.2
L3	Connect [L1, L2, L3 of E_c, L3 of E_s, L1, L2 of FF]

Table 4: The network architecture of FF module.

Layer	Dec
L1	Upsample:2, SPADE, Resnet
L2	Upsample:2, SPADE, Resnet
L3	SPADE, Resnet
L4	Conv(I:64,O:3,K:7,P:3,S:1), tanh

Table 5: The network architecture of Encoder Dec.

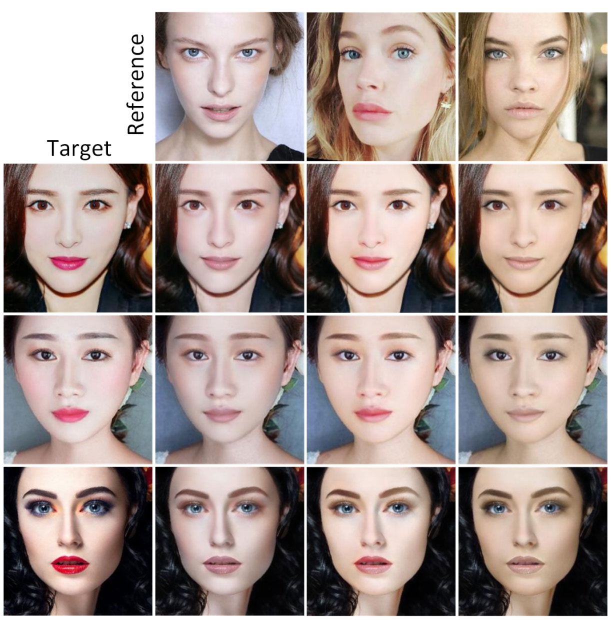

Figure 13: Our makeup removal results. The first column is three makeup target images, the first row is three non-makeup reference images, the makeup removal results are displayed in the lower right corner.

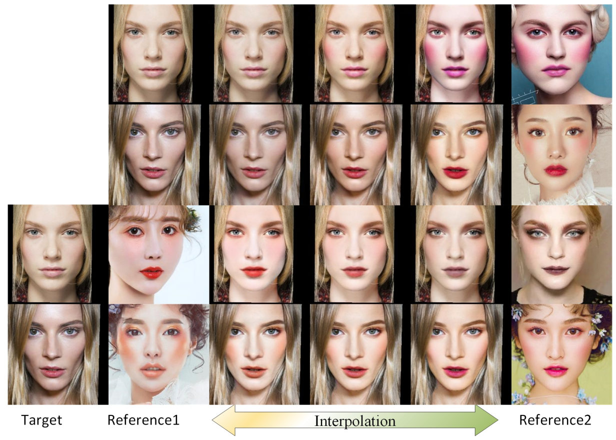

Figure 14: Results of makeup style interpolation. The first four rows have one reference image, and the last four rows have two. The makeup styles of the interpolated images continuously transfer from reference 1 to reference 2.

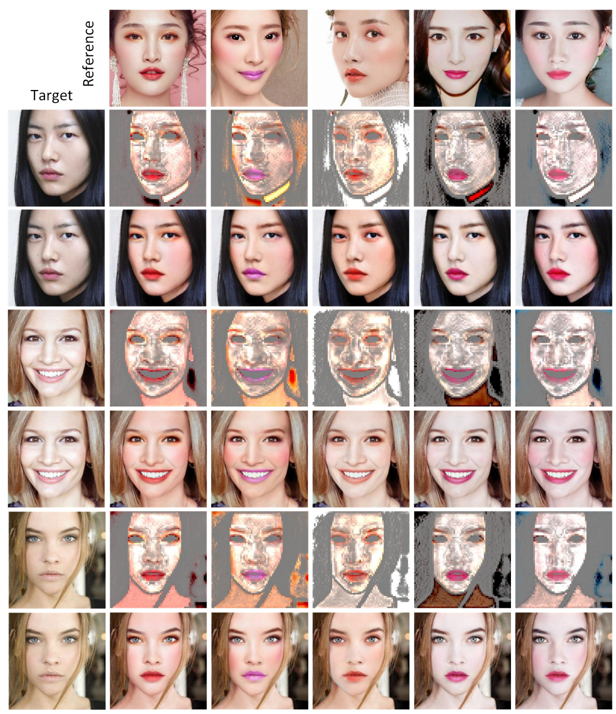

Figure 15: More of our makeup transfer results. The first column is three target images, the first row is five reference images, the semantic correspondence results and the makeup transfer results are displayed in the lower right corner..

References

Chang, H.; Lu, J.; Yu, F.; and Finkelstein, A. 2018. PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup. In *Computer Vision and Pattern Recognition*, 40–48.

Chen, H.-J.; Hui, K.-M.; Wang, S.-Y.; Tsao, L.-W.; Shuai, H.-H.; and Cheng, W.-H. 2019. BeautyGlow: On-Demand Makeup Transfer Framework With Reversible Generative Network. In *Computer Vision and Pattern Recognition*.

Deng, H.; Han, C.; Cai, H.; Han, G.; and He, S. 2021. Spatially-Invariant Style-Codes Controlled Makeup Transfer. In *Computer Vision and Pattern Recognition*, 6549–6557.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, 2672–2680.

Gu, Q.; Wang, G.; Chiu, M. T.; Tai, Y.-W.; and Tang, C.-K. 2019. LADN: Local Adversarial Disentangling Network for Facial Makeup and De-Makeup. In *International Conference on Computer Vision*, 10481–10490.

Guo, D.; and Sim, T. 2009. Digital face makeup by example. In *Computer Vision and Pattern Recognition*, 73–79.

He, M.; Chen, D.; Liao, J.; Sander, P. V.; and Yuan, L. 2018. Deep Exemplar-based Colorization. In *International Conference on Computer Graphics and Interactive Techniques*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal Unsupervised Image-to-Image Translation. In *European Conference on Computer Vision*, 179–196.

Huang, Z.; Zheng, Z.; Yan, C.; Xie, H.; Sun, Y.; Wang, J.; and Zhang, J. 2020. Real-World Automatic Makeup via Identity Preservation Makeup Net. In *International Joint Conference on Artificial Intelligence*.

Jiang, W.; Liu, S.; Gao, C.; Cao, J.; He, R.; Feng, J.; and Yan, S. 2020. PSGAN: Pose and Expression Robust Spatial-Aware GAN for Customizable Makeup Transfer. In *Computer Vision and Pattern Recognition*, 5194–5202.

Karras, T.; Laine, S.; and Aila, T. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Computer Vision and Pattern Recognition*, 4401–4410.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. ImageNet classification with deep convolutional neural networks. *Communications of The ACM*, 60: 84–90.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M. K.; and Yang, M.-H. 2018. Diverse Image-to-Image Translation via Disentangled Representations. In *European Conference on Computer Vision*, 36–52.

Lee, J.; Kim, E.; Lee, Y.; Kim, D.; Chang, J.; and Choo, J. 2020. Reference-Based Sketch Image Colorization Using Augmented-Self Reference and Dense Semantic Correspondence. In *Computer Vision and Pattern Recognition*, 5800–5809.

Li, C.; Zhou, K.; and Lin, S. 2015. Simulating makeup through physics-based manipulation of intrinsic image layers. In *Computer Vision and Pattern Recognition*, 4621–4629.

Li, T.; Qian, R.; Dong, C.; Liu, S.; Yan, Q.; Zhu, W.; and Lin, L. 2018. BeautyGAN: Instance-level Facial Makeup Transfer with Deep Generative Adversarial Network. In *ACM Multimedia*, 645–653.

Liao, J.; Yao, Y.; Yuan, L.; Hua, G.; and Kang, S. B. 2017. Visual attribute transfer through deep image analogy. In *ACM Transactions on Graphics*, volume 36.

Liu, S.; Jiang, W.; Gao, C.; He, R.; Feng, J.; Li, B.; and Yan, S. 2021. PSGAN++: Robust Detail-Preserving Makeup Transfer and Removal. In *ArXiv, abs/2105.12324*.

Lyu, Y.; Dong, J.; Peng, B.; Wang, W.; and Tan, T. 2021. SOGAN: 3D-Aware Shadow and Occlusion Robust GAN for Makeup Transfer. In *ArXiv, abs/2104.10567*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y. K.; Wang, Z.; and Smolley, S. P. 2017. Least Squares Generative Adversarial Networks. In *International Conference on Computer Vision*, 2813–2821.

Nguyen, T.; Tran, A. T.; and Hoai, M. 2021. Lipstick Ain't Enough: Beyond Color Matching for In-the-Wild Makeup Transfer. In *Computer Vision and Pattern Recognition*, 13305–13314.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Computer Vision and Pattern Recognition*, 2337–2346.

Sarfraz, M. S.; Seibold, C.; Khalid, H.; and Stiefelhagen, R. 2019. Content and Colour Distillation for Learning Image Translations with the Spatial Profile Loss. In *British Machine Vision Conference*, 287.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Sun, Z.; Liu, F.; Liu, R. W.; Xiong, S.; and Liu, W. 2020. Local Facial Makeup Transfer via Disentangled Representation. In *Asian Conference on Computer Vision*.

Tong, W.-S.; Tang, C.-K.; Brown, M. S.; and Xu, Y.-Q. 2007. Example-Based Cosmetic Transfer. In *Pacific Conference on Computer Graphics and Applications*, 211–218.

Wan, Z.; Chen, H.; Zhang, J.; Jiang, W.; Yao, C.; and Luo, J. 2021. Facial Attribute Transformers for Precise and Robust Makeup Transfer. In *ArXiv, abs/2104.02894*.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In *European Conference on Computer Vision*, 334–349.

Zhang, P.; Zhang, B.; Chen, D.; Yuan, L.; and Wen, F. 2020. Cross-Domain Correspondence Learning for Exemplar-Based Image Translation. In *Computer Vision and Pattern Recognition*, 5143–5153.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *International Conference on Computer Vision*.