SLGAN: Style- and Latent-guided Generative Adversarial Network for Desirable Makeup Transfer and Removal

Daichi Horita
The University of Tokyo
horita@hal.t.u-tokyo.ac.jp

Kiyoharu Aizawa
The University of Tokyo
aizawa@hal.t.u-tokyo.ac.jp

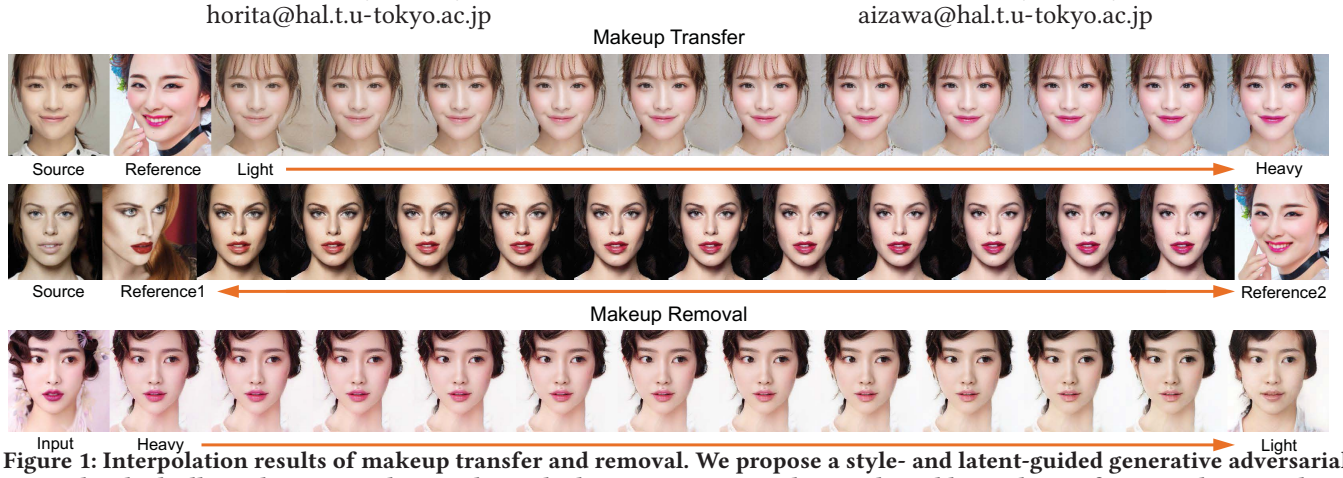

Figure 1: Interpolation results of makeup transfer and removal. We propose a style- and latent-guided generative adversarial network, which allows the user to adjust makeup shading in an image to obtain a desirable result. Our framework interpolates from light to heavy makeup based on a style-guided value with a single reference image (first row) and two reference images (second row). Our framework can also arbitrarily remove makeup by modulating a latent-guided value (third row).

ABSTRACT

There are five features to consider when using generative adversarial networks to apply makeup to photos of the human face. These features include (1) facial components, (2) interactive color adjustments, (3) makeup variations, (4) robustness to poses and expressions, and the (5) use of multiple reference images. To tackle the key features, we propose a novel style- and latent-guided makeup generative adversarial network for makeup transfer and removal. We provide a novel, perceptual makeup loss and a style-invariant decoder that can transfer makeup styles based on histogram matching to avoid the identity-shift problem. In our experiments, we show that our SLGAN is better than or comparable to state-of-the-art methods. Furthermore, we show that our proposal can interpolate facial makeup images to determine the unique features, compare existing methods, and help users find desirable makeup configurations.

CCS CONCEPTS

• **Computing methodologies** → **Image representations**.

KEYWORDS

GANs, image translation, makeup transfer, makeup removal

ACM Reference Format:
Daichi Horita and Kiyoharu Aizawa. 2022. SLGAN: Style- and Latent-guided Generative Adversarial Network for Desirable Makeup Transfer and Removal. In *ACM Multimedia Asia (MMAsia '22), December 13–16, 2022, Tokyo, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3551626.3564967

MMAsia '22, December 13–16, 2022, Tokyo, Japan

ACM ISBN 978-1-4503-9478-9/22/12.
https://doi.org/10.1145/3551626.3564967

1 INTRODUCTION

Virtual makeup applications help us to produce makeup interactively using pre-defined filters. In reality, mastering makeup is not easy because it needs expertise in cosmetic products and techniques. It remains challenging to produce makeup as the user desire with the virtual makeup tools.

We consider five features important when virtually making up our faces in photos. These features are (1) facial components, (2) interactive color adjustments, (3) makeup variations, (4) robustness to poses and expressions, and (5) the use of multiple reference images. Several studies of makeup transfer (MT) and makeup removal (MR) have been proposed [1, 2, 8, 10]. However, current works do not satisfy all five mentioned features.

In this paper, we propose a style- and latent-guided GAN (SLGAN). As shown in Figures 1 and 4, our framework effectively performs MT and MR while accounting for the five features mentioned above. Our framework employs style- and latent-guided translation for the tasks. As shown in Figure 1, our framework can interpolate makeup shading. Thus, users can adjust the generated results to find desirable makeup combinations with their reference images. Furthermore, as shown in Figure 4, our method is robust to poses and expressions.

Moreover, we propose a novel perceptual makeup loss to help the generator apply more appropriate makeup features to the input image. The loss function is used to compute a histogram matching between the generated image and the reference image using the features extracted by a style encoder. It thus enables our framework to adequately transfer makeup styles.

The major contributions of this paper are summarized as follows:

(1) We propose a SLGAN framework for an MT and MR. This is the first style- and latent-guided framework for this task.

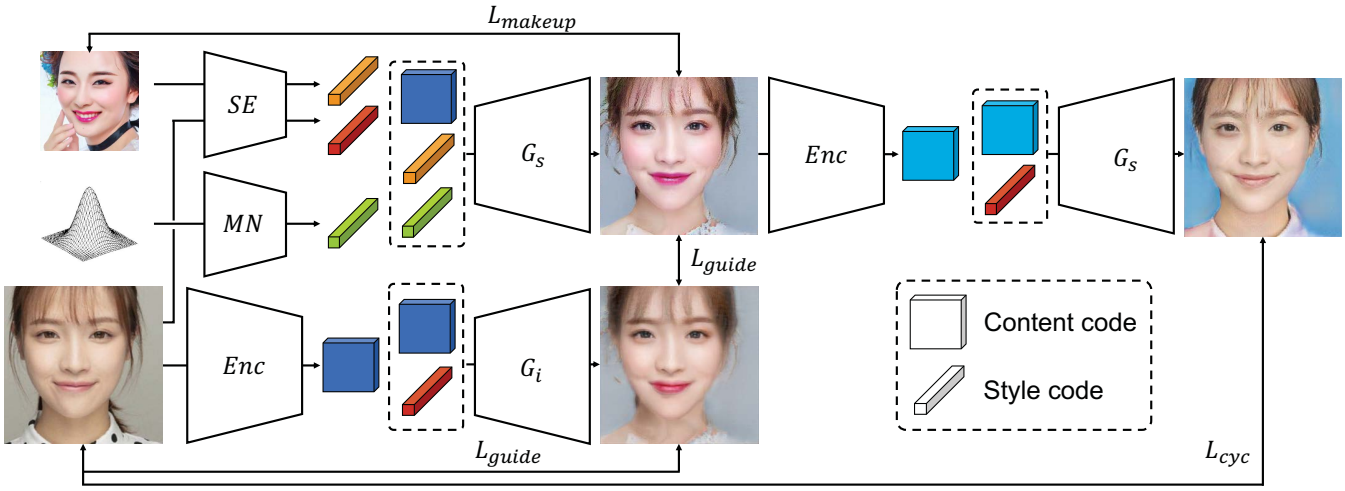

Figure 2: SLGAN consists of four modules: a generator G, a style encoder E_S, and a mapping network E_M. The generator G consists of a shared encoder G_E, a style-guided decoder G_S, and a style-invariant decoder G_I.

(2) Our proposed style-invariant decoder assists the generator to translate images that preserve the identity of the source.
(3) We propose a novel perceptual makeup loss that enables the generator to perform a high-quality translation.
(4) Quantitative and qualitative experimental results show that SLGAN is better than or comparable to state-of-the-art methods.

2 RELATED WORKS

2.1 Unpaired Image-to-Image Translation

In an unpaired image-to-image translation, the problem is translating an input image into its corresponding output image. StarGANv2 [4] provided both latent- and reference-guided synthesis. However, StarGANv2 causes an identity-shift problem, which is the generated image loses the contents of the source image because they embed global style features of the reference image. For example, when a person with long hair is given as a reference, an output image with long hair is always generated. Thus, these methods are not suitable for MT and MR problems. To overcome this problem, our method embeds references in style codes considering local features.

2.2 Makeup Studies

MT is to perform style transfer considering the semantics of a source and reference image. BeautyGAN [10] simultaneously trained MT and MR using a single generator and discriminator. Additionally, BeautyGAN proposed a makeup loss, which matched the color histogram between the generated and reference images of facial components (e.g., lips, eye shadows, and whole face). To effectively utilize the loss for our framework, we propose a perceptual makeup loss that optimizes the network to embed a reference image in a style code based on makeup styles.

LADN [5] proposed local discriminators, which learned the makeup features of each facial component. PSGAN [8] performs an MT using an attention mechanism based on semantic information of facial landmarks and masks with a style-guided architecture. However, this method fails to generate images when this additional semantic information is not obtained. In contrast, we propose a framework that does not depend on such information.

3 SLGAN

3.1 Formulation

Our goal is to extract makeup styles from the reference images and transfer them to the source images. We have I_s^X to represent source samples, I_r^Y to represent reference samples, and z to represent latent codes. We utilize one-hot vectors c to represent makeup conditions.

3.2 Network Architecture

Overall As shown in Figure 2, we propose a style- and latent-guided framework for MT and MR. First, the style codes s_e, s_m are generated by a style encoder SE and a non-linear mapping network MN from the reference image $I_r^Y \in Y$ and the latent code $z \in Z$, respectively. Then, given an embedded style code s and a source image I_s^X, our goal is to learn a generator $G : I_s^X, s \rightarrow \tilde{I}_r^X$.

Style encoder. Given a reference image I_r^Y and a one-hot vector c, the style encoder SE learns to embed the reference image into a style code s_e, denoted as $SE_c : I_r^Y, c \rightarrow s_e \in W$. SE extracts the feature using an encoder and then applies the MLP layers per domain based on the control of the one-hot vector c. Therefore, because the style encoder uses each MLP layer for MT and MR, it can embed reference images in style codes with domain-specific representations.

Mapping network. Given a latent code z in the input latent space Z and a random one-hot vector c, our non-linear mapping network MN learns to embed a latent code in the style code s_m, denoted as $MN_c : z, c \rightarrow s_m \in W$. Our mapping network also yields a domain-specific style code s_m.

Adaptive normalization layer. We use AdaIN [6] with the style-guided decoder to perform MT and MR based on the style codes s_e and s_m of I_r^Y and z, respectively. The style codes s_e and s_m control β and γ in the AdaIN operation after each convolution layer of the generator G. The features of each source image I_s^X are individually normalized and the scaling and shifting operations are performed using scalar components based on the style codes s.

Generator. Given a source image I_s^X and a style code s, our generator generates an image $\tilde{I}_r^X$, that preserves both the makeup style of the reference image I_r^Y, and the identity of the source I_s^X. As shown in Figure 2, our generator G consists of an encoder Enc, a style-guided decoder G_s, and a style-invariant decoder G_i.

To simplify the notation, we denote a style-guided generator as $G_{sg}(I_s^X, s) = G_s(Enc(I_s^X), s)$ and a style-invariant generator as $G_{ig}(I_s^X) = G_i(Enc(I_s^X))$. Each decoder has the same structure except for the normalization layer. An encoder Enc embeds a source image I_s^X in a content code. An encoder and a style-invariant decoder have an instance normalization so that they can make the features conform to a normal distribution.

Discriminator. To make the generator G generate realistic images, we use the discriminator D. Our discriminator D has the same structure as the style encoder SE.

3.3 Style-invariant Decoder

There is an identity-shift problem in which the generator cannot preserve the identity of the source image when the global style of the reference image is embedded in the style code. Thus, the discrepancy of identities between reference and source images can cause problems in which the generated image cannot maintain the content of the source. To overcome this problem, our proposed style-invariant decoder generates images from the shared feature without the style code, which is extracted by a shared encoder. That is, this network has no AdaIN layers. On the other hand, our style-invariant decoder helps the generator not only perform stable learning like the guide decoder but also helps avoid an identity-shift problem.

3.4 Perceptual Makeup Loss

To further encourage the network to transfer makeup per face component, we propose a new histogram-matching strategy and propose perceptual makeup loss. Our key idea is for the style encoder to have a structure for extracting makeup and non-makeup styles. The perceptual makeup loss computes the histogram matching using features of each convolution layer of a style encoder between the generated image and the reference image. This encourages the style encoder to learn better parameters through multi-task learning. The loss entails the integration of three local histogram losses acting on the lips, eyes, and facial regions, defined as

$$\mathcal{L}_{makeup} = \lambda_{lips}\mathcal{L}_{lips} + \lambda_{eyes}\mathcal{L}_{eyes} + \lambda_{face}\mathcal{L}_{face}, \tag{1}$$

$$\mathcal{L}_{item} = \sum_{l=1}^{K} ||\phi_l(\tilde{I}_r^X) - HM(\phi_l(\tilde{I}_r^X \circ S_{item}^1), \phi_l(I_r^Y \circ S_{item}^2))||_2, \tag{2}$$

$$S_{item}^1 = FP(\tilde{I}_r^X), S_{item}^2 = FP(I_r^Y), \tag{3}$$

where ϕ_l denotes a l-th layer feature map, K denotes the sum of the number of convolution layers, $\circ$ denotes element-wise multiplication, $item$ denotes the set of $\{lips, eyes, face\}$, FP denotes the face parsing algorithm, HM denotes the histogram matching operation, and S denotes the semantic mask of face components.

3.5 Other Objectives

Additionally, regarding the perceptual makeup loss described in Section 3.4, we use the following objectives, which are similar to related works [4, 12–14].

Adversarial Loss. To make the generated images more realistic, we adopt an adversarial loss, defined as

$$\begin{aligned}\mathcal{L}_{adv} = \min_{G_s}\max_{D_c} &\mathbb{E}_{I_s^X, c}\left[\log D_c(I_s^X)\right] + \\ &\mathbb{E}_{I_s^X, \hat{c}, \hat{s}}[\log(1 - D_{\hat{c}}(G_{sg}(I_s^X, \hat{s})))],\end{aligned} \tag{4}$$

where the target style code $\hat{s}$ is generated by a style encoder $\hat{s_e} = SE_{\hat{c}}(I_r^Y)$ and a non-linear mapping network $\hat{s_m} = MN_{\hat{c}}(z/)$. c and $\hat{c}$ represent the source domain and target domain, respectively. D_c represents the corresponding domain of c and G_{sg} represents the style-guided generator. A discriminator distinguishes whether the generated image $\tilde{I}_r^X$ is a real or not.

Style diversity loss. We introduce a regularization term to spread over the generated space [12, 13], which is defined as

$$\mathcal{L}_{sd} = \mathbb{E}_{I_s^X, \hat{c}, z_1, z_2}\left[||G_{sg}(I_s^X, \hat{s_1})) - G_{sg}(I_s^X, \hat{s_2})||_1\right], \tag{5}$$

where $\hat{s_1}$ and $\hat{s_2}$ are generated by a style encoder $SE_{\hat{c}}$ or a mapping network $MN_{\hat{c}}$ from random latent codes z_1 and z_2, and a target condition vector $\hat{c}$, denoted as $\hat{s_e} = SE_{\hat{c}}(z)$ and $\hat{s_m} = MN_{\hat{c}}(z)$, respectively. This encourages the generator to explore the latent code and increases the chance of generating various samples. The discriminator learns better parameters, because it properly classifies samples that are rarely generated. As a result, by using this objective, our framework properly learns fine makeup styles.

Style reconstruction loss. To constrain the style codes to correctly represent the style of makeup or non-makeup, we use the style reconstruction loss [7, 15], defined as

$$\mathcal{L}_{sr} = \mathbb{E}_{I_s^X, \hat{c}, z}\left[||\hat{s} - SE_{\hat{c}}(G_{sg}(I_s^X, \hat{s}))||_1\right]. \tag{6}$$

This objective is similar to a latent reconstruction loss [3, 4].

Cycle consistency loss. By optimizing Eq.(4,5,6), the generator can generate diverse and realistic images. However, the generator should not only preserve the features of the source image, but it should also fool the discriminator. As a result, there is a problem in which only these objectives do not guarantee that the generated image preserves the content of the source image. To solve this problem, we use the cycle consistency loss [9, 14], defined as

$$\mathcal{L}_{cyc} = \mathbb{E}_{I_s^X, c, \hat{c}, s}\left[||I_s^X - G_{sg}(G_{sg}(I_s^X, \hat{s}), \bar{s})||_1\right], \tag{7}$$

where $\hat{s}$ represents the style code of a target domain $\hat{c}$ and $\bar{s}$ represents an original domain c of I_s^X, denoted as $\bar{s} = SE_c(I_s^X)$. Minimizing this objective enables the generator to perform a MR and MR while preserving the contents of the source image.

Style-invariant guide loss. Despite the use of cycle consistency loss, the generated image changes the shape of facial components, depending on makeup and non-makeup styles, owing the identity-shift problem. To achieve this problem, we propose a style-invariant guide loss to encourage the generated image to naturally apply the style of the reference image and maintain the content of the source image. Is is defined as

$$\begin{aligned}\mathcal{L}_{guide} = &\mathbb{E}_{I_s^X}\left[\lambda_\gamma||I_s^X - G_{ig}(I_s^X)||_2\right] + \\ &\mathbb{E}_{I_s^X, \hat{c}, s}\left[\lambda_\beta||G_{ig}(I_s^X) - G_{sg}(I_s^X, \hat{s})||_2\right],\end{aligned} \tag{8}$$

where each λ are hyper-parameters, G_{ig} represents the style-invariant generator, G_{sg} represents the style-guided generator, and $\hat{s}$ represents the style code of the target domain $\hat{c}$. We do not give the style code $\hat{s}$ to the style-invariant generator G_{ig}.

Total Loss. Finally, the loss functions of G, SE, MN, and D, which are optimized in our framework, are defined as

$$\mathcal{L}_D = -\mathcal{L}_{adv} \tag{9}$$

$$\mathcal{L}_G = \mathcal{L}_{adv} - \mathcal{L}_{sd} + \mathcal{L}_{sr} + \mathcal{L}_{cyc} + \mathcal{L}_{makeup} + \mathcal{L}_{guide}. \tag{10}$$

4 EXPERIMENTS

4.1 Implementation Details

We use the Makeup Transfer (MT) dataset [10] for MT and MR. The dataset contains 3,834 facial images with a resolution of 256 × 256, consisting of 1,115 non-makeup images and 2,719 makeup unpaired images. The test set consists of 100 and 250 non-makeup and makeup images, respectively.

Figure 3: Qualitative comparison of makeup transfer of baseline methods and our style-guided SLGAN. Our method can generate images that are closer to the reference image from the views of lips, eyes, eye shadows, and skin tones.

Figure 4: Qualitative comparison of makeup transfer with the source and reference images having different poses.

4.2 Makeup Transfer Results

We compared our style-guided SLGAN with baseline models on an MT task. As baseline methods, we adopted two general image-to-image translation methods: DIA [11] and CycleGAN [14]. We adopted five MT methods: BeautyGAN [10], PairedCycleGAN [1], BeautyGlow [2], LADN [5], and PSGAN [8].

Qualitative Comparison. Figure 3 shows qualitative comparisons of SLGAN with the baseline methods. DIA and CycleGAN failed to transfer makeup for the eyebrows and lip color, respectively. In the lower row, BeautyGlow's eye shadow was clearly darker than that of the reference image. LADN generated artifacts around the hair. In the lower image, PSAGN failed to transfer the eye color and eye shadow. The other baseline methods failed to transfer the pupil color of the reference image. We argue that these capabilities are important because people often use colored contact lenses to change their eye colors. Compared with the baseline methods, our style-guided SLGAN generated images that have reference makeup.

Figure 4 shows the results of different poses of source and reference images. BeautyGAN and LADN failed to transfer makeup or generate artifacts. These methods did not provide an explicit structure for learning MT locations, and they overfitted the MT dataset, which contained only frontal images. In contrast, SLGAN succeeded in transferring makeup. Our framework learned the relationships between each face part because the perceptual makeup loss was computed between the features of our style encoder.

Quantitative Comparison. We conducted a user study using Amazon Mechanical Turk (AMT) in which 10 people participated. Given each generated image, a corresponding source, and a reference image, the Turkers were instructed to choose a natural image following a reference makeup. From the generated results, we randomly selected 50 images per method. Table 1 shows the results of the 10-person user study. In this small-scale experiment, our style-guided SLGAN had a better score, compared with the other methods.

4.3 Makeup Removal Results

Qualitative Comparison. Figure 5 shows makeup removal results. CycleGAN showed a blurred image of poor quality. PairedCycleGAN and LADN tended to remove makeup, however, they failed to generate clear lips and eyes. In contrast, we found that our method produced clear MR images. We observed that the images generated by our style-guided SLGAN were affected by the skin color of the given reference image.

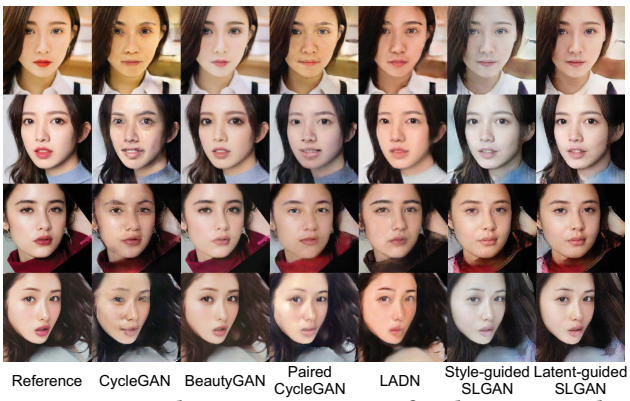

Figure 5: Qualitative comparison of makeup removal.

Table 1: AMT evaluation for a makeup transfer and removal.

Method	Transfer (%) ↑	Removal (%) ↑
CycleGAN	17.2	5.6
PairedCycleGAN	19.2	13.2
BeautyGAN	22.0	1.8
LADN	18.4	0.4
Style-guided SLGAN	**23.2**	38.6
Latent-guided SLGAN	-	**40.4**

Qualitative Comparison. We also conducted a user study using AMT as same as MT. Table 1 shows both our style- and latent-guided SLGAN showed better results compared with the baseline methods. We can see that our style- and latent-guided SLGAN demonstrated similar quality MR with few differences. We consider that our style-guided SLGAN performed MR based upon the skin color of the reference image, and it, therefore, scored lower than our latent-guided SLGAN.

5 CONCLUSION

We have presented SLGAN, which is a style- and latent-guided framework for MT and MR. Our perceptual makeup loss enables our framework to adequately transfer makeup styles. Our style-invariant decoder further enabled our framework to avoid the identity-shifting problem. In the experiments, our SLGAN performed better or comparably to state-of-the-art methods, and the unique ability to interpolate the MT and MR results.

REFERENCES

[1] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. 2018. PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup. In *CVPR*.

[2] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. 2019. BeautyGlow: On-Demand Makeup Transfer Framework With Reversible Generative Network. In *CVPR*.

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. arXiv:1606.03657

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *CVPR*.

[5] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. 2019. LADN: Local Adversarial Disentangling Network for Facial Makeup and De-Makeup. In *ICCV*.

[6] Xun Huang and Serge Belongie. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In *ICCV*.

[7] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-Image Translation. In *ECCV*.

[8] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. 2020. PSGAN: Pose and Expression Robust Spatial-Aware GAN for Customizable Makeup Transfer. In *CVPR*.

[9] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. 2018. Diverse Image-to-Image Translation via Disentangled Representations. In *ECCV*.

[10] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. 2018. BeautyGAN: Instance-Level Facial Makeup Transfer with Deep Generative Adversarial Network. In *ACMMM*.

[11] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual Attribute Transfer through Deep Image Analogy. *ACM Trans. Graph.* 36, 4 (2017).

[12] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. 2019. Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis. In *CVPR*.

[13] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. 2019. Diversity-Sensitive Conditional Generative Adversarial Networks. In *ICLR*.

[14] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *ICCV*.

[15] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. 2017. Toward Multimodal Image-to-Image Translation. In *NeurIPS*.