BeautyGAN: Instance-level Facial Makeup Transfer with Deep Generative Adversarial Network

Tingting Li*
Tsinghua-Berkeley Shenzhen Institute, Tsinghua University
litt.thu@foxmail.com

Ruihe Qian
Institue of Information Engineering of CAS
vitochien09@gmail.com

Chao Dong†
SIAT-Sensetime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
chao.dong@siat.ac.cn

Si Liu
Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University
liusi@buaa.edu.cn

Qiong Yan
SenseTime Research
yanqiong@sensetime.com

Wenwu Zhu
Tsinghua-Berkeley Shenzhen Institute, Department of Computer Science and Technology, Tsinghua University
wwzhu@tsinghua.edu.cn

Liang Lin
Sun Yat-sen University
linliang@ieee.org

ABSTRACT

Facial makeup transfer aims to translate the makeup style from a given reference makeup face image to another non-makeup one while preserving face identity. Such an instance-level transfer problem is more challenging than conventional domain-level transfer tasks, especially when paired data is unavailable. Makeup style is also different from global styles (e.g., paintings) in that it consists of several local styles/cosmetics, including eye shadow, lipstick, foundation, and so on. Extracting and transferring such local and delicate makeup information is infeasible for existing style transfer methods. We address the issue by incorporating both global domain-level loss and local instance-level loss in an dual input/output Generative Adversarial Network, called BeautyGAN. Specifically, the domain-level transfer is ensured by discriminators that distinguish generated images from domains' real samples. The instance-level loss is calculated by pixel-level histogram loss on separate local facial regions. We further introduce perceptual loss and cycle consistency loss to generate high quality faces and preserve identity. The overall objective function enables the network to learn translation on instance-level through unsupervised adversarial learning. We also build up a new makeup dataset that consists of 3834 high-resolution face images. Extensive experiments show that BeautyGAN could generate visually pleasant makeup faces and accurate transferring results. Data and code are available at http://liusi-group.com/projects/BeautyGAN.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**; **Neural networks**; *Unsupervised learning*;

KEYWORDS

facial makeup transfer; generative adversarial network

ACM Reference Format:
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 *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240618

1 INTRODUCTION

Makeup is a ubiquitous way to improve one's facial appearance with special cosmetics, such as foundation for concealing facial flaws, eye liner, eye shadow and lipstick. With thousands of cosmetic products, varying from brands, colors, way-to-use, it is difficult to find a well-suited makeup style without professional suggestions. Virtual makeup application is a convenient tool, helping users try the makeup style from photos, such as MEITU XIUXIU, TAAZ and DailyMakever[1]. However, these tools all require user's manual interaction and provide with only a certain number of fixed makeup styles. In daily life, celebrities always wear beautiful makeup styles, which give some examples to refer to. Makeup transfer provides an efficient way to help users select the most suitable makeup style. As shown in Figure 1, makeup transfer (results of our method) could

*This work was done when Tingting Li and Ruihe Qian were interns at Sensetime.
†corresponding author

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240618

[1]xiuxiu.web.meitu.com, taaz.com, dailymakeover.com

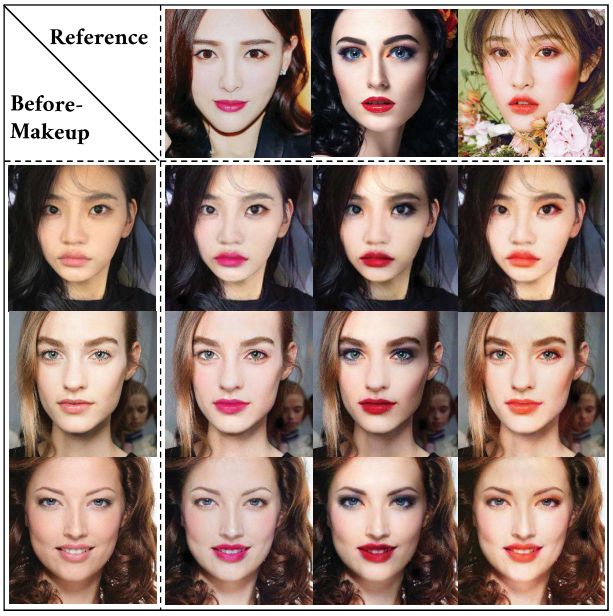

Figure 1: Example results of our BeautyGAN model for makeup transfer. Three makeup styles on reference images (top row) are translated to three before-makeup images (left column). Nine generated images are shown in the middle.

translate the makeup style from a given reference face image to another non-makeup face without the change of face identity.

Existing studies on automatic makeup transfer can be classified into two categories: traditional image processing approaches[11, 19, 29] like image gradient editing and physics-based manipulation, and deep learning based methods which typically build upon deep neural networks[23]. Image processing approaches generally decompose images into several layers (e.g., face structure, color, skin) and transfer each layer after warping the reference face image to the non-makeup one. Deep learning based method[23] adopt serveral independent networks to deal with each cosmetic individually. Almost all previous methods treat the makeup style as a simple combination of different components, thus the overall output image looks unnatural with apparent artifacts at combining places (see Figure 4).

Recent progress on image-to-image translation, such as style transfer[8, 9, 13], has shown that an end-to-end structure act on the entire image could generate high quality results. However, directly applying these techniques in our task is still infeasible. Facial makeup transfer has two main characteristics that are different from previous problems. 1) Makeup style varies from face to face, and require transferring on instance-level. On the contrary, typical image-to-image translation methods[4, 12, 35] built upon generative adversarial networks (GAN) are mostly for domain-level transfer. For instance, CycleGAN[35] realizes image-to-image translation between two collections (e.g., horses and zebras), and emphasizes inter-domain differences while omits intra-domain differences. Therefore, using CycleGAN in our problem tends to generate an average domain-level style that is invariant given different reference faces (see Figure 4). 2) Makeup style is beyond a global style and includes independent local styles. Specifically, in conventional style transfer works[8, 9, 13], style generally refers to the global painting manner like brush stroke and color distribution. In contrast, makeup style is more delicate and elaborate, which consists of several local cosmetics including eye shadows, lipsticks, foundation and so on. Each cosmetic represents a completely different style. Therefore, it is difficult to extract makeup style as a whole while preserving particular traits of various cosmetics.

Another crucial issue is the lack of training data. On one hand, the released makeup dataset (see Table 1) is too small to train a sufficient large network, and the facial makeup images are mostly of low resolution/quality. On the other hand, it is difficult to obtain a pair of well-aligned face images with different makeup styles. Thus supervised learning with paired data is also implausible.

To address the above issues, we propose a novel dual input/output generative adversarial network called BeautyGAN, to realize makeup style transfer in an unified framework. It accepts the makeup and non-makeup faces as inputs and directly outputs the transferred results. No additional pre-/post-processing is required. Similar to CycleGAN[35], the network first transfers the non-makeup face to the makeup domain with a couple of discriminators that distinguish generated images from domains' real samples. On the basis of domain-level transfer, we achieve instance-level transfer by adopting a pixel-level histogram loss calculated on different facial regions. To preserve face identity and eliminate artifacts, we also incorporate a perceptual loss and a cycle consistency loss in the overall objective function. Thanks to the dual input/output design, the cycle consistency between inputs and outputs could be achieved with only one generator, which realizes makeup and anti-makeup simultaneously in a single forward pass. Moreover, no paired data is needed during the whole training procedure. As shown in Figure 1, the generated images are natural-looking and visually pleasant without observable artifacts.

To sum up, the main contributions are three-folds:

(1) We achieve automatic makeup transfer with a dual input/output generative adversarial network. Experiments present the effectiveness of the transferring strategy, and generated results are of higher quality than state-of-the-art methods.
(2) We achieve instance-level style transfer by successfully applying pixel-level histogram losses on local regions. Such instance-level transfer approach can be easily generalized to other image translation tasks, such as style transfer for head-shot portraits, image attribute transfer and so on.
(3) we build up a new makeup dataset with a collection of 3834 images, which is available at http://liusi-group.com/projects/BeautyGAN.

2 RELATED WORKS

2.1 Makeup Studies

Recently, makeup related studies have aroused much more attention. [31] proposed a facial makeup detector and remover framework based on locality-constrained dictionary learning. [20] introduced an adversarial network to generate non-makeup images for makeup-invariant face verification. Makeup transfer is another attractive

application, which aims to transfer makeup style from reference image when still preserving source image identity. [11] decomposed images into three layers and transferred makeup information layer by layer. This method may smooth facial details of source images thus another image decomposition method was introduced by [19]. All above makeup transfer frameworks are based on traditional methods, while [23] proposed a localized makeup transfer framework in the way of deep learning. It divided facial makeup into several parts and conducted different methods on each facial part. Warping and structure preservation were employed to synthesize after-makeup images.

Unlike the aforementioned works, our network could realize makeup transfer and makeup removal simultaneously. Meanwhile, the unified training process could consider relationships among cosmetics in different regions. In addition, the end-to-end network itself could learn the adaptation of cosmetics fed in source images, thus eliminates the need of post-processing.

2.2 Style Transfer

Style transfer aims to combine content and style from different images. To achieve this goal, [8] proposed a method that generated a reconstruction image by minimizing the content and the style reconstruction loss. To control more information like color, scale and spatial location, an improved approach was presented in [9], where perceptual factors were introduced. The methods mentioned above could produce high-quality results but require heavy computation. [13] proposed a feed-forward network for style transfer with less computation and approximate quality.

2.3 Generative Adversarial Networks

Generative Adversarial Networks[10] (GANs) is one of the generative models, consisting a discriminator and a generator. GAN has been widely used in computer vision tasks due to its ability of generating visually realistic images. [17] presented a generative adversarial network for image super resolution. [6] employed conditional GAN[25] to solve particular eye in-painting problem. [27] trained adversarial models on synthetic images for improving the realism of them. [34] even enabled to incorporate user interactions to present real-time image editing, where GAN was leveraged to estimate the image manifold.

2.4 GAN for Image-to-Image Translation

Most existing researches on image-to-image translation aim to learn a mapping from source domain to target domain. Recently, there are some promising works[4, 12, 35] appling GAN to this field. [12] proposed a so-called pix2pix framework, which could sythesize images from label maps and reconstruct objects from edge images. To solve the problem of lacking paired images for training, [22] proposed a model whose generators were bounded with weight-sharing constraints to learn a joint distribution. [35][14] presented cycle consistency loss to regularize the key attributes between inputs and translated images. StarGAN[4] even solved problem of mapping among multiple domains within one single generator. Specially, [15] introduced an encoder working with GAN for image attribute transfer.

3 OUR APPROACH: BEAUTYGAN

Our goal is to realize facial makeup transfer between a reference makeup image and a non-makeup image on instance-level. Consider two data collections, $A \subset \mathbb{R}^{H\times W\times 3}$ referring to non-makeup image domain and $B \subset \mathbb{R}^{H\times W\times 3}$ referring to makeup image domain with various makeup styles on. We simultaneously learn the mapping between two domains, denoted as $G : A \times B \rightarrow B \times A$, where '$\times$' represents Cartesian product. That is to say, given two images as inputs: a source image $I_{src} \in A$ and a reference image $I_{ref} \in B$, the network is expected to generate an after-makeup image $I^B_{src} \in B$ and an anti-makeup image $I^A_{ref} \in A$, denoted as $(I^B_{src}, I^A_{ref}) = G(I_{src}, I_{ref})$. I^B_{src} synthesizes the makeup style of I_{ref} while preserving the face identity of I_{src}, and I^A_{ref} realizes makeup removal from I_{ref}. The fundamental problem is how to learn instance-level correspondence, which should ensure the makeup style consistency between result image I^B_{src} and reference images I_{ref}. Note that there is no paired data for training.

To address the issue, we introduce pixel-level histogram loss 像素级别的直方图损失 acted on different cosmetics. In addition, perceptual loss has been 感知损失 employed to maintain face identity and structure. Then we can transfer the exact makeup to the source image without the change of face structure. The proposed method is based on Generative Adversarial Networks[10], and it is convenient to integrate all loss 使结合 terms into one full objective function. Adversarial losses help generate visually pleasant images and refine the correlation among different cosmetics. The details of loss functions and network architectures are shown below.

3.1 Full Objective

As illustrated in Figure 2, the overall framework consists of one generator G and two discriminators: D_A, D_B. In the formulation $(I^B_{src}, I^A_{ref}) = G(I_{src}, I_{ref})$, G accepts two images, $I_{src} \in A$ and $I_{ref} \in B$, as inputs and generates two translated images as outputs, $I^B_{src} \in B$ and $I^A_{ref} \in A$.

We first give objective functions of D_A and D_B, which contain only adversarial losses. D_A distinguishes the generated image I^A_{ref} from real samples in set A, given by:

$$\begin{aligned}\mathcal{L}_{D_A} &= \mathbb{E}_{I_{src}}[logD_A(I_{src})] \\ &+ \mathbb{E}_{I_{src},I_{ref}}[log(1 - D_A(I^A_{ref}))].\end{aligned} \tag{1}$$

Similarly, D_B aims to distinguish generated image I^B_{src} from real samples in set B, given by:

$$\begin{aligned}\mathcal{L}_{D_B} &= \mathbb{E}_{I_{ref}}[logD_A(I_{ref})] \\ &+ \mathbb{E}_{I_{src},I_{ref}}[log(1 - D_A(I^B_{src}))].\end{aligned} \tag{2}$$

The full objective function of generator G contains four types of losses: adversarial loss, cycle consistency loss, perceptual loss and makeup constrain loss,

$$\mathcal{L}_G = \alpha\mathcal{L}_{adv} + \beta\mathcal{L}_{cyc} + \gamma\mathcal{L}_{per} + \mathcal{L}_{makeup}, \tag{3}$$

where α, β, γ are weighting factors that controls the relative importance of each term. Adversarial loss for G integrates two terms: L_{D_A} and L_{D_B} as 集成

$$\mathcal{L}_{adv} = \mathcal{L}_{D_A} + \mathcal{L}_{D_B}. \tag{4}$$

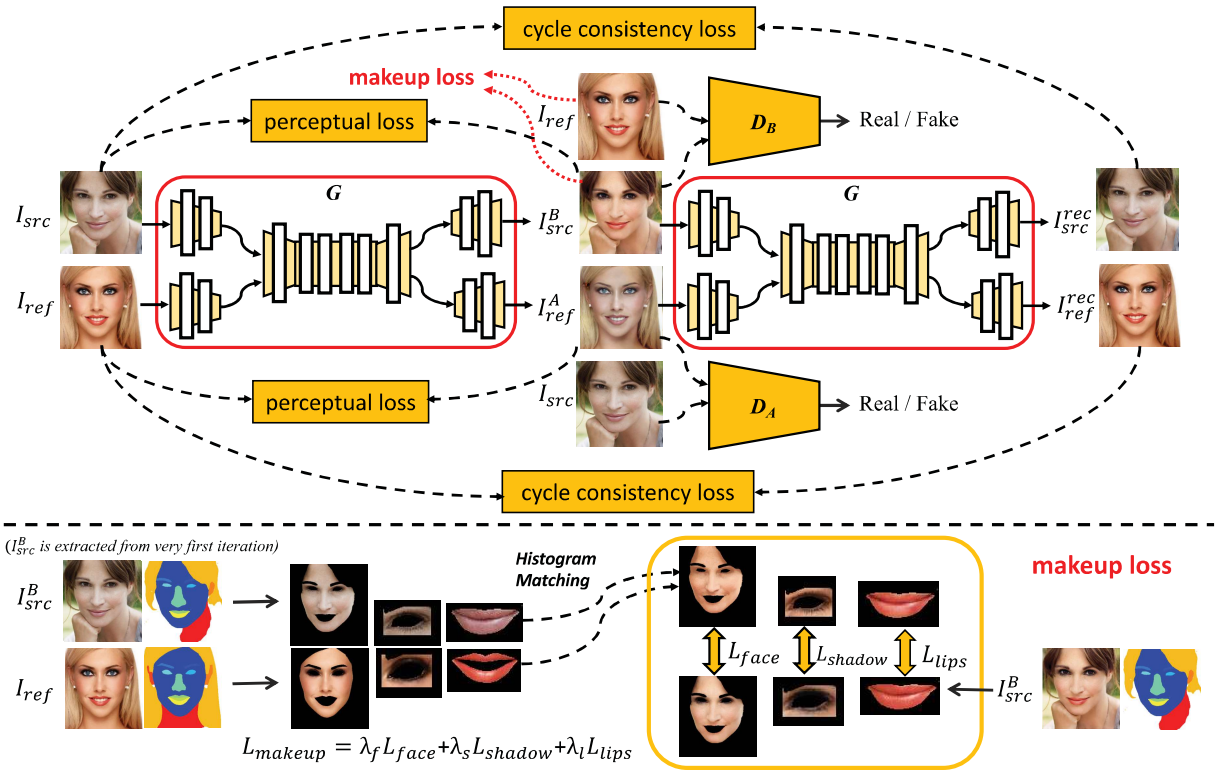

Figure 2: Framework of the proposed BeautyGAN. The upper pipeline shows the overall system. G accpets two images as inputs: non-makeup image I_{src}, reference makeup image I_{ref}, and generates two outputs: transferred makeup image I^B_{src}, anti-makeup image I^A_{ref}. The generated images are fed into the same G to build up reconstruction results: $I^{rec}_{src}, I^{rec}_{ref}$. There are four loss terms for training G: cycle consistency loss, perceptual loss, adversarial loss (denoted as D_A and D_B) and makeup loss. The lower pipeline shows the details of makeup loss. It consists of three local histogram loss terms acted on face, eye shadow and lips, respectively. We first utilize face parsing model to separate each cosmetic region of $I_{src}, I_{ref}, I^B_{src}$. Then, for each region, we employ histogram matching between I_{src} and I_{ref} to obtain a histogram remapping facial region as ground truth. The local loss term calculates pixel-level differences between such ground truth and corresponding cosmetic region of I^B_{src}.

Note that the generator G and discriminators D_A, D_B play minmax game, where G tries to minimize the adversarial loss and discriminators D_A, D_B aim to maximize the same loss function. Three remaining losses will be detailed in the subsequent sections.

3.2 Domain-Level Makeup Transfer

We exploit domain-level makeup transfer as the foundation of instance-level makeup transfer. Thanks to the dual input/output architecture, the proposed network could simultaneously learn the 同时 mapping between two domains within just one generator. The output images are required to preserve face identities and background information as input images. To satisfy these two constraints, we impose perceptual loss and cycle consistency loss, respectively.

Rather than directly measuring differences between pixel-level Euclidean distance, perceptual loss calculates differences between high-level features extracted by deep convolutional networks. In this paper, we utilize 16-layer *VGG* networks pre-trained on ImageNet Dataset. For an image x, $F_l(x)$ denotes the corresponding feature maps in lth layer on *VGG*, where $F_l \in \mathbb{R}^{C_l \times H_l \times W_l}$. C_l is the number of feature maps, H_l and W_l are height and width of each feature map, respectively. Thus the perceptual loss between input images I_{src}, I_{ref} and output images I^B_{src}, I^A_{ref} are expressed as:

$$\mathcal{L}_{per} = \frac{1}{C_l \times H_l \times W_l} \sum_{ijk} E_l \tag{5}$$

$$E_l = [F_l(I_{src}) - F_l(I^B_{src})]^2_{ijk} + [F_l(I_{ref}) - F_l(I^A_{ref})]^2_{ijk}, \tag{6}$$

where F^l_{ijk} is the activation of the ith filter at position $<j, k>$ in lth layer.

In order to maintain background information, we also introduce cycle consistency loss. When the output images are passed into the generator, it is supposed to generate images as close as the original input images. This procedure can be expressed as

$$(I_{src}, I_{ref}) \rightarrow G(I_{src}, I_{ref}) \rightarrow G(G(I_{src,I_{ref}})) \approx (I_{src}, I_{ref}). \tag{7}$$

The loss function is formulated as

$$\mathcal{L}_{cyc} = \mathbb{E}_{I_{src},I_{ref}}[dist(I^{rec}_{src}, I_{src}) + dist(I^{rec}_{ref}, I_{ref})], \quad (8)$$

where $(I^{rec}_{src}, I^{rec}_{ref}) = G(G(I_{src}, I_{ref}))$. The distance function $dist(\cdot)$ could be chosen as $L1$ norm, $L2$ norm or other metrics.

3.3 Instance-level Makeup Transfer

To further encourage the network to learn instance-level makeup transfer, it is essential to add constraints on makeup style consistency. We observe that facial makeup could be visually recognized as color distributions. No matter lipsticks, eye shadows or foundations, the makeup process could be mainly understood as color changing. There are various color transfer methods that can be found in the survey [7]. We employ Histogram Matching (HM), a straightforward method, and introduce additional histogram loss on pixel-level, which restricts the output image I^B_{src} and the reference image I_{ref} to be identical in makeup style.

Histogram loss. If we directly adopt MSE loss on pixel-level histograms of two images, the gradient will be zero, owning to the indicator function, thus makes no contribution to optimization process. Therefore, we adopt histogram matching strategy that generates a ground truth remapping image in advance. Suppose that we want to calculate histogram loss on pixels between original image x and reference image y, we should first perform histogram matching on x and y to obtain a remapping image $HM(x, y)$, which has the same color distribution as y but still preserves content information as x. After we get $HM(x, y)$, it is convenient to utilize the MSE loss between $HM(x, y)$ and x, then back-propagate the gradients for optimization.

Face parsing. Instead of utilizing histogram loss over the entire image, we split the makeup style into three important components – lipsticks, eye shadow, foundation. State-of-the-art methods like [23] also take these three components as makeup representations. And then we apply localized histogram loss on each part. The reasons are two folds. First, pixels in background and hairs have no relationship with makeup. If we do not separate them apart, they will disturb the correct color distribution. Second, facial makeup is beyond a global style but a collection of several independent styles in different cosmetics regions. In that sense, we employ the face parsing model in[32] to obtain face guidance mask as $M = FP(x)$. For each input image x, pre-trained face parsing model would generate an index mask M denoting several facial locations, including lips, eyes, face skin (corresponds to foundation), hairs, background and so on. At last, for each M, we track different labels to produce three corresponding binary masks, representing for cosmetics spatiality: $M_{lip}, M_{eye}, M_{face}$. It is important to note that eye shadows are not annotated on M, because the before-makeup images have no eye shadows. But we expect the result image I^B_{src} to have similar eye shadow color and shape as reference image I_{ref}. According to eyes mask M_{eye}, we calculate two rectangle areas enclosing eye shadows and then exclude eyes regions, some hair and eyebrow regions in between. Thus we could create a specific binary mask representing for eye-shadows M_{shadow}.

计算眼影周围的矩形区域，并排除眼部区域，以及头发和眉毛

Makeup loss. The overall makeup loss are integrated by three local histogram losses acted on lips, eye shadows and face regions, respectively:

$$\mathcal{L}_{makeup} = \lambda_l \mathcal{L}_{lips} + \lambda_s \mathcal{L}_{shadow} + \lambda_f \mathcal{L}_{face}, \quad (9)$$

where $\lambda_l, \lambda_s, \lambda_f$ are weight factors. We multiply images with their corresponding binary masks and process spatially histogram matching between result image I^B_{src} and reference image I_{ref}. Formally, we define local histogram loss as

$$\mathcal{L}_{item} = \|I^B_{src} - HM(I^B_{src} \circ M^1_{item}, I_{ref} \circ M^2_{item})\|_2. \quad (10)$$

$$M^1 = FP(I^B_{src}) \quad (11)$$

$$M^2 = FP(I_{ref}) \quad (12)$$

Here, $\circ$ denotes element-wise multiplication and *item* are in set of $\{lips, shadow, face\}$.

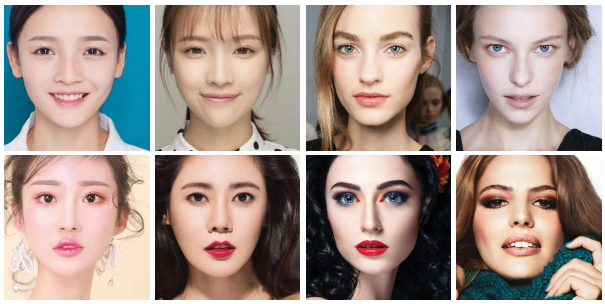

Figure 3: Samples from MT dataset. The non-makeup and makeup images are shown in upper row and lower row.

Table 1: Comparison with released makeup datasets.

Dataset	Subjects	Images per subject	Total number of images
YMU[2]	151	4	604
VMU[5]	51	4	204
MIW[1]	125	1-2	154
MIFS[3]	214	4 or 2	642
Ours(MT)	3000+	1-2	3834

4 DATA COLLECTION

We collect a new facial makeup dataset consisting of 3834 female images in total, with 1115 *non-makeup* images and 2719 *makeup* images. We refer to this dataset as the Makeup Transfer(MT) dataset. It includes some variations in race, pose, expression and background clutter. Plenty of makeup styles have been assembled, including smoky-eyes makeup style, flashy makeup style, Retro makeup style, Korean makeup style and Japanese makeup style, varying from subtle to heavy. Specifically, there are some nude makeup images, as for convenience, have been classified into non-makeup category.

The initial data are crawled from websites. We manually remove low resolution images under bad illumination condition. And then retained images are employed face alignment with 68 landmarks. According to the two eye locations, we transform them to the same

spatial size 256 × 256. Among 3834 images, we randomly select 100 non-makeup images and 250 makeup images for test. The remaining images are separated into training set and validation set.

MT is the biggest makeup dataset comparing to other released makeup datasets. Existing makeup datasets mostly consist of no more than 1000 images. They typically assemble pairs of images for one subject: before-makeup image and after-makeup image pair. Although the same object is obtained in such pair images, they have differences on view: poses, expressions even illumination. Generally, they are applied for studying the impact of makeup in face recognition and are inapplicable for makeup transfer task. MT dataset contains diverse makeup styles and more than 3000 subjects. The detailed comparison between makeup datasets are listed in table 1. Examples of MT are illustrated in Figure 3.

5 EXPERIMENTS

In this section, we depict the network architecture and training setting. All the experiments apply the same MT dataset we release. We compare performances of our method and some other baselines from both qualitative and quantitative perspectives. And we further give a thorough analysis on components of BeautyGAN.

5.1 Implementation Details

Network architecture. We design the generator G with two inputs and two outputs. To be specific, the network contains two separate input branches with convolutions, respectively. In the middle we concatenate these two branches together and feed them into several residual blocks. After that, the output feature maps will be upsampled by two individual branches of transposed convolutions to generate two result images. Note that the branches do not share parameters within layers. We also use instance normalization [30] for G. As for discriminators D_A and D_B, we leverage identical 70×70 PatchGANs [18], which classifies local overlapping image patches to be real or fake.

Training Details. To stabilize the training procedure and generate high quality images, we apply two additional training strategies. First, inspired by [35], we replace all negative log likelihood objective in adversarial loss by a least square loss [24]. For instance, equation 1 is then defined as below, so as equation 2 and 4:

$$\begin{aligned}\mathcal{L}_{D_A} &= \mathbb{E}_{I_{src}}[(D_A(I_{src})-1)^2] \\ &+ \mathbb{E}_{I_{src},I_{ref}}[D_A(I^A_{ref})^2].\end{aligned} \tag{13}$$

Second, we introduce spectral normalization [26] for stably training discriminators. It is computationally light and easy to incorporate, which satisfies the Lipschitz constraint $\sigma(W) = 1$:

$$W_{SN}(W) := \frac{W}{\sigma(W)}, \tag{14}$$

where $\sigma(W)$ is the spectral norm of W, denoted as:

$$\sigma(w) := \max_{h:h\neq 0} \frac{\|Wh\|_2}{\|h\|_2} = \max_{\|h\|_2\leq 1} \|Wh\|_2, \tag{15}$$

here, h is the input of each layer.

For all experiments, we obtain masks annotated labels on different facial regions through a PSPNet[33] trained for face segmentation. The *relu_4_1* feature layer of *VGG16*[28] network is applied for identity preserving. Such *VGG16* is pre-trained on ImageNet and has parameters fixed all through training process. The parameters in equations 3 and 9 are: $\alpha = 1, \beta = 10, \gamma = 0.005, \lambda_l = 1, \lambda_s = 1, \lambda_f = 0.1$. We train the network from scratch using Adam[16] with learning rate of 0.0002 and batch size of 1.

5.2 Baselines

Digital Face Makeup[11] is an early makeup transfer work, applying traditional image processing method.
DTN[23] is the state-of-the-art makeup transfer work. It proposes a deep localized makeup transfer network, which independently transfers different cosmetics.
Deep Image Analogy [21] is a recent work, which realizes visual attribute transfer across two semantic-related images. It adapts image analogy to match features extracted from deep neural networks. We apply it on makeup transfer task for comparison.
CycleGAN[35] is an representative unsupervised image-to-image translation work. To adapt makeup transfer task, we modify the generator in it with two branches as input, but maintain all the other architecture and setting as the same.
Style Transfer[13] is a related work, which trains a feed-forward network for synthesizing style and content information from respective images. We employ non-makeup image as content and reference makeup image as style for experiments.

5.3 Comparison Against Baselines

Qualitative evaluation. As demonstrated in Figure 4, we show qualitative comparison results with baselines. We observe that although Guo *et al.* [11] produces images with makeup on, the results all have visible artifacts. It seems like a fake mask attached on the non-makeup face. Mismatch problem occurs around facial and eyes contour. Some incorrect details are transferred, such as the black eye shadows on the second and fourth rows are transferred into blue. Liu *et al.* [23] transfers different cosmetics independently, as a result, it shows alignment artifacts around eye areas and lips area. The foundation and eye shadows have not been correctly transferred as well. Style transfer[13] generates images introducing grain-like artifacts, which deteriorate image quality. It typically transfers global style like painting strokes thus is infeasible for delicate makeup style transfer. Comparing to the above methods, CycleGAN[35] could produces relatively realistic images. However, the makeup styles are not consistent with references. Liao *et al.* [21] produces outputs with similar makeup styles as references and shows natural results. However, it transfers not only facial makeup, but also other features in reference images. For example, the third image changes background color from black to gray, the fourth image changes hair color and all images modify pupil colors to be like references. In addition, it transfers lighter makeup styles than references, especially in lipsticks and eye shadows.

Compared with baselines, our method generates high quality images with the most accurate makeup styles on, no matter compared in eye shadows, lipsticks or foundations. For instance, in the second row, only our result transfers the dark eye shadows of reference image. The results also show that BeautyGAN keep other makeup-irrelevant components intact as the original non-makeup images, like hairs, clothes and background. In Figure 5, we zoom

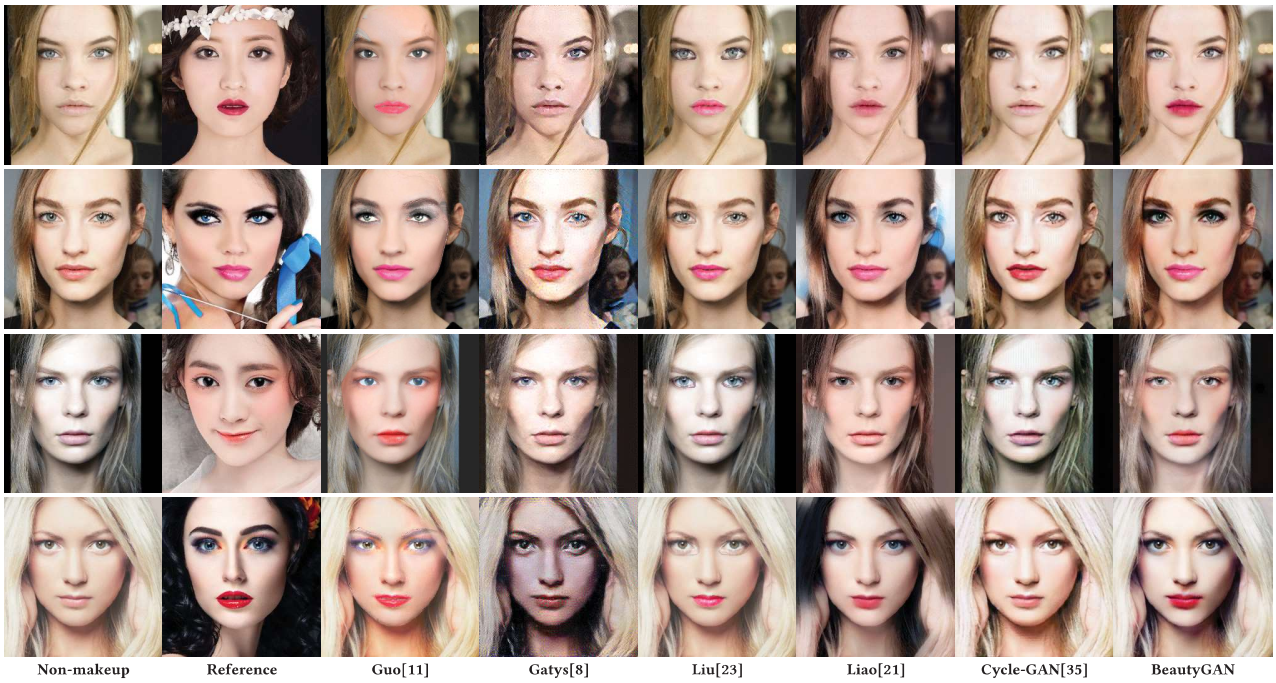

Figure 4: Qualitative comparisons between BeautyGAN and baselines. The first two columns are non-makeup images and reference images, respectively. Each row in the remaining columns shows the makeup transfer results of different methods.

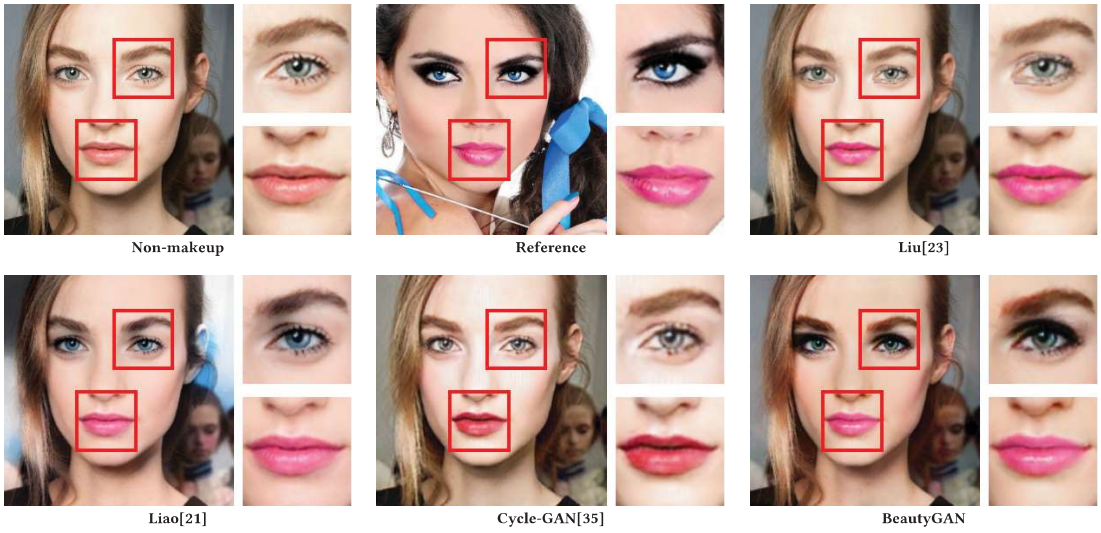

Figure 5: Zoom in the performance of eye makeup and lipsticks transfer.

in the performances of eye makeup and lipsticks transfer for better demonstrating the comparison. More results are shown in the supplementary file.

Quantitative comparison. For quantitative evaluation on BeautyGAN, we conduct a user study from 84 volunteers. We randomly choose 10 non-makeup test images and 20 makeup test images, which would obtain 10×20 after-makeup results for each makeup transfer method. Two representative baselines are in comparison: Liao *et al.* [21] and Liu *et al.* [23]. Each time, we present five images, including a non-makeup image, a makeup image as reference, and three randomly shuffled makeup transfer images generated from different methods. Participants are instructed to give a rank order of

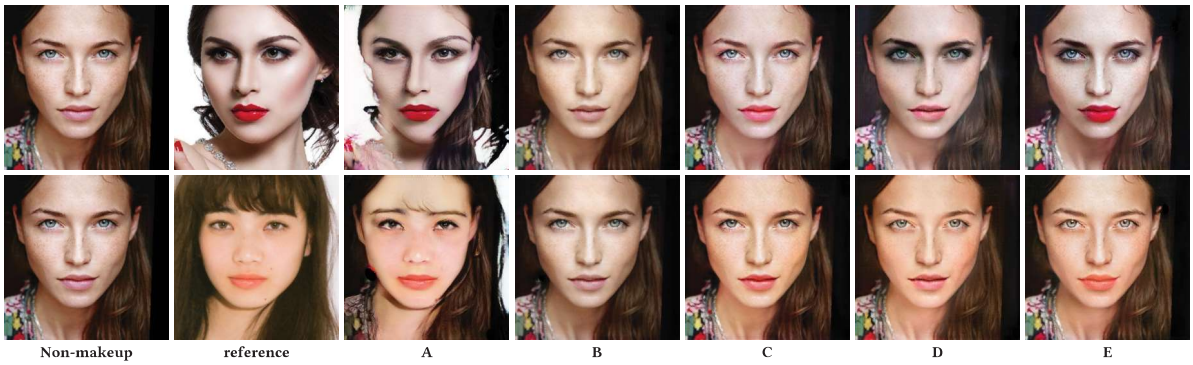

Figure 6: Results of ablation study. The first two columns are non-makeup images and reference images, respectively. Each row in the remaining columns shows the makeup transfer results of five experiments: A, B, C, D, E. Different experiment settings are illustrated in Table 2.

Table 2: Ablation Study Setups

Setup	$\mathcal{L}_{per}$	$\mathcal{L}_{face}$	$\mathcal{L}_{eye-shadow}$	$\mathcal{L}_{lips}$
A		√	√	√
B	√			
C	√	√		
D	√	√	√	
E	√	√	√	√

Table 3: Result of user study.

Methods	Rank 1	Rank 2	Rank 3
Liu[23]	4.25%	10.78%	84.97%
Liao[21]	33.91%	46.03%	20.06%
Ours	61.84%	27.56%	10.59%

three generated images, based on quality, realism and makeup style similarity. Rank 1 represents the best makeup transfer performance while rank 3 represent the worst makeup transfer performance.

Table 3 shows the results. For each method, we normalize the votes and obtain the percentages of three rank orders. There are 61.84% results of BeautyGAN rank number one, comparing to 33.91% of Liao *et al.* and 4.25% of Liu *et al.*. Also, BeautyGAN has the least percentage on Rank 3 column. We observe that BeautyGAN is mostly voted as Rank 1, Liao *et al.* distributes mainly on Rank 2 and Liu *et al.* has most votes on Rank 3. User study demonstrates that our method performs better than other baselines.

5.4 Component Analysis of BeautyGAN

To investigate the importance of each component in overall objective function (Eqn. 3), we perform ablation studies. We mainly analyse the effect of perceptual loss term (Eqn. 5) and makeup loss term (Eqn. 9). Thus the experiments are conducted with adversarial and cycle consistency loss all the time. Table 2 shows the settings and Figure 6 demonstrates the results.

In experiment A, we remove the perceptual loss term from Eqn. 3. In such situation, the results are all fake images like two inputs warped and merged on pixels. On the contrary, other experiments, where perceptual loss term is included, show that the identities of non-makeup faces are maintained. Therefore, it indicates perceptual loss helps to preserve image identity.

Experiments B, C, D, E are designed for investigating makeup loss term, which consists of three local histogram loss acted on different cosmetic regions: $\mathcal{L}_{face}, \mathcal{L}_{shadow}, \mathcal{L}_{lips}$. In experiment B, we directly remove makeup loss from Eqn. 3. We find the generated images are slightly modified on skin tone and lipsticks, but do not transfer makeup style from reference images. We then sequentially add $\mathcal{L}_{face}$, $\mathcal{L}_{shadow}$ and $\mathcal{L}_{lips}$ on experiment C, D, E. Column C shows results that foundation could be successfully transferred from reference images. Based on foundation transfer, experiment D add eye shadow constraints $\mathcal{L}_{shadow}$ within. We observe that local eye makeup are transferred as well. Column E is the results trained with overall makeup loss. It shows that lipsticks are additionally transferred comparing to column D. To sum up, makeup loss is necessary for instance-level makeup transfer. Three terms of makeup loss play role on foundation, eye shadow and lipsticks transfer, respectively.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a dual input/output BeautyGAN for instance-level facial makeup transfer. With one generator, BeautyGAN could realize makeup and anti-makeup simultaneously in a single forward pass. We introduce pixel-level histogram loss to constrain the similarity of makeup style. Perceptual loss and cycle consistency loss have been employed to preserve identity. Experimental results demonstrate that our approach can achieve significant performance gain over existing approaches.

Acknowledgments: This work was supported by the National Program on Key Basic Research Project (No. 2015CB352300), National Natural Science Foundation of China Major Project (No. U1611461), National Natural Science Foundation of China (Nos. U1536203, 61572493), IIE project (No. Y6Z0021102, No. Y7Z0241102), CCF-Tencent Open Research Fund and International Partnership Program of Chinese Academy of Sciences (172644KYSB20160033).

REFERENCES

[1] Cunjian Chen, Antitza Dantcheva, and Arun Ross. 2013. Automatic facial makeup detection with application in face recognition. In *Biometrics (ICB), 2013 International Conference on*. IEEE, 1–8.

[2] Cunjian Chen, Antitza Dantcheva, and Arun Ross. 2016. An ensemble of patch-based subspaces for makeup-robust face recognition. *Information fusion* 32 (2016), 80–92.

[3] Cunjian Chen, Antitza Dantcheva, Thomas Swearingen, and Arun Ross. 2017. Spoofing faces using makeup: An investigative study. In *Identity, Security and Behavior Analysis (ISBA), 2017 IEEE International Conference on*. IEEE, 1–8.

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2017. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *arXiv preprint arXiv:1711.09020* (2017).

[5] Antitza Dantcheva, Cunjian Chen, and Arun Ross. 2012. Can facial cosmetics affect the matching accuracy of face recognition systems?. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*. IEEE, 391–398.

[6] Brian Dolhansky and Cristian Canton Ferrer. 2017. Eye In-Painting with Exemplar Generative Adversarial Networks. *arXiv preprint arXiv:1712.03999* (2017).

[7] Hasan Sheikh Faridul, Tania Pouli, Christel Chamaret, Jürgen Stauder, Alain Trémeau, Erik Reinhard, et al. 2014. A Survey of Color Mapping and its Applications. *Eurographics (State of the Art Reports)* 3 (2014).

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[9] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. 2017. Controlling perceptual factors in neural style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

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

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2016. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004* (2016).

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*. Springer, 694–711.

[14] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192* (2017).

[15] Taeksoo Kim, Byoungjip Kim, Moonsu Cha, and Jiwon Kim. 2017. Unsupervised visual attribute transfer with reconfigurable generative adversarial networks. *arXiv preprint arXiv:1707.09798* (2017).

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2016. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802* (2016).

[18] Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*. Springer, 702–716.

[19] Chen Li, Kun Zhou, and Stephen Lin. 2015. Simulating makeup through physics-based manipulation of intrinsic image layers. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4621–4629.

[20] Yi Li, Lingxiao Song, Xiang Wu, Ran He, and Tieniu Tan. 2017. Anti-Makeup: Learning A Bi-Level Adversarial Network for Makeup-Invariant Face Verification. *arXiv preprint arXiv:1709.03654* (2017).

[21] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 120.

[22] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *Advances in neural information processing systems*. 469–477.

[23] Si Liu, Xinyu Ou, Ruihe Qian, Wei Wang, and Xiaochun Cao. 2016. Makeup like a superstar: deep localized makeup transfer network. In *the Association for the Advance of Artificial Intelligence*. AAAI Press, 2568–2575.

[24] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, and Zhen Wang. 2016. Multi-class Generative Adversarial Networks with the L2 Loss Function. *CoRR* abs/1611.04076 (2016). arXiv:1611.04076 http://arxiv.org/abs/1611.04076

[25] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[26] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* (2018).

[27] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. 2016. Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv:1612.07828* (2016).

[28] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

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

[30] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2016. Instance Normalization: The Missing Ingredient for Fast Stylization. *CoRR* abs/1607.08022 (2016). arXiv:1607.08022 http://arxiv.org/abs/1607.08022

[31] Shuyang Wang and Yun Fu. 2016. Face Behind Makeup. In *the Association for the Advance of Artificial Intelligence*. 58–64.

[32] Zhen Wei, Yao Sun, Jinqiao Wang, Hanjiang Lai, and Si Liu. 2017. Learning Adaptive Receptive Fields for Deep Image Parsing Network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2434–2442.

[33] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2881–2890.

[34] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. 2016. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*. Springer, 597–613.

[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593* (2017).