LADN: Local Adversarial Disentangling Network for Facial Makeup and De-Makeup

Qiao Gu*
CMU, HKUST
qiaog@andrew.cmu.edu

Guanzhi Wang*
Stanford University, HKUST
guanzhi@stanford.edu

Mang Tik Chiu
UIUC
mtchiu2@illinois.edu

Yu-Wing Tai
Tencent
yuwingtai@tencent.com

Chi-Keung Tang
HKUST
cktang@cs.ust.hk

Abstract

We propose a local adversarial disentangling network (LADN) for facial makeup and de-makeup. Central to our method are multiple and overlapping local adversarial discriminators in a content-style disentangling network for achieving local detail transfer between facial images, with the use of asymmetric loss functions for dramatic makeup styles with high-frequency details. Existing techniques do not demonstrate or fail to transfer high-frequency details in a global adversarial setting, or train a single local discriminator only to ensure image structure consistency and thus work only for relatively simple styles. Unlike others, our proposed local adversarial discriminators can distinguish whether the generated local image details are consistent with the corresponding regions in the given reference image in cross-image style transfer in an unsupervised setting. Incorporating these technical contributions, we achieve not only state-of-the-art results on conventional styles but also novel results involving complex and dramatic styles with high-frequency details covering large areas across multiple facial features. A carefully designed dataset of unpaired before and after makeup images is released at https://georgegu1997.github.io/LADN-project-page.

1. Introduction

We propose to incorporate local adversarial discriminators into an image domain translation network for details transfer between two images, and apply these local adversarial discriminators on overlapping image regions to achieve image-based facial makeup and removal. By encouraging cross-cycle consistency between input and output, we can disentangle the makeup latent variable from other factors on a single facial image. Through increasing the number and overlapping local discriminators, complex makeup styles with high-frequency details can be seamlessly transferred or removed while facial identity and structure are both preserved. See Figure 1.

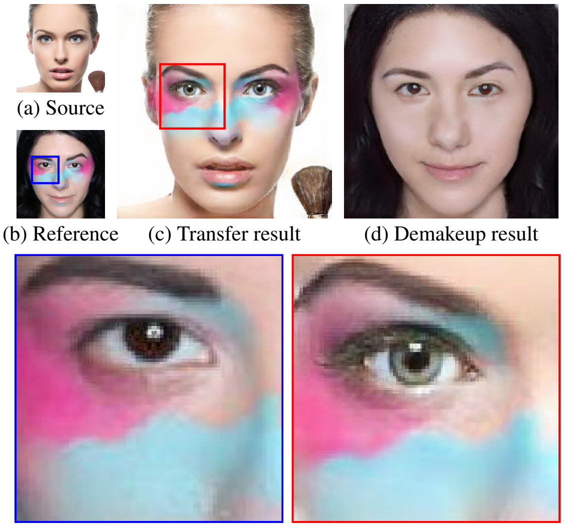

(a) Source
(b) Reference
(c) Transfer result
(d) Demakeup result
(e) Zoom-in view

Figure 1: Facial makeup and de-makeup with dramatic makeup style. See supplemental material for high-quality images of all our results.

The contributions of our paper are:

- By utilizing *local adversarial discriminators* rather than cropping the image into different local paths, our network can seamlessly transfer and remove dramatic makeup styles;
- Through incorporating *asymmetric loss functions* on makeup transfer and removal branches, the network

*Equal contribution. Authorship order was determined by rolling dice.

This research is supported in part by Tencent and the Research Grant Council of the Hong Kong SAR under grant no. 1620818.

is forced to disentangle the makeup latent variable from others, and thus our network can generate photorealistic results where facial identity is mostly preserved;

- A dataset containing unpaired before-makeup and after-makeup facial images will be released for non-commercial purpose upon the paper's acceptance.

Our target application, digital facial makeup [2, 1], has been increasingly popular. Its inverse application, known as facial de-makeup [27, 8] is also starting to gain more attention. All current results in deep learning only either work for or demonstrate conventional or relatively simple makeup styles, possibly due to limitations of their network architectures and overfitting to their datasets. Existing methods often fail to transfer/remove dramatic makeup, which oftentimes is the main usage for such an application, before the user physically applies the dramatic makeup which may take hours to accomplish.

Given an image of a clean face without makeup as the source, and another image of an after-makeup face as the reference, the makeup transfer problem is to synthesize a new image where the specific makeup style from the reference is applied on the face of the source (Figure 1). The main problem stems from the difficulty of extracting the makeup-only latent variable, which is required to be disentangled from other factors in a given facial image. This problem is often referred to as content-style separation. Most existing works addressed this problem through region-specific style transfer and rendering [22, 7, 21, 23, 3]. This approach can precisely extract the makeup style in specific and well-defined facial regions such as eyes and mouth where makeup is normally applied, but it limits the application range in the vicinity of these facial regions, and thus fails to transfer/remove more dramatic makeup where color and texture details can be far away from these facial features.

By incorporating multiple and overlapping local discriminators in a content-style disentangling network, we successfully perform transfer (resp. removal) of complex/dramatic makeup styles with all details faithfully transferred (resp. removed).

2. Related Work

Given the bulk of deep learning work on photographic image synthesis, we will review related work in image translation and style transfer, and those on makeup transfer. We will also review approaches that involve global and local discriminators and describe the differences between ours and theirs.

Style transfer and image domain translation. Style transfer can be formulated as an image domain translation problem, which was first formulated by Taigman *et al.* [25] as learning a generative function to map a sample image from a source domain to a target domain. Isola *et al.* [13] proposed the pix2pix framework which adopted a conditional GAN to model the generative function. This method, however, requires cross-domain, paired image data for training. Zhu *et al.* [28] introduced the CycleGAN to relax this paired data requirement, by incorporating a cycle consistency loss into the generative network to generate images that satisfy the distribution of desired domain. Lee *et al.* [15] recently proposed a disentangled representation framework, DRIT, to diversify the outputs with unpaired training data by adding a reference image from the target domain as input. They encode images into a domain-invariant content space and another domain-specific attribute space. By disentangling content and attribute, the generated output adopts the content of an image in another domain while preserving the attributes of its own domain. However, in the context of makeup/de-makeup transfer, DRIT can only be applied when the relevant makeup style transfer can be formulated into image-to-image translation. As our experiments show, this means that only light makeup styles can be handled.

Method	Work
Global discriminator	GAN [9], pix2pix[13]
Single local discriminator	Image completion[12, 19], PatchGAN [16], CycleGAN[28]
Multiple overlapping local discriminators	**LADN** (ours)

Table 1: Related works on local and global discriminators. Different from existing works, our paper applies multiple local discriminators in overlapping image regions.

Makeup transfer and removal. Tong *et al.* [26] first tackled this problem by solving the mapping of cosmetic contributions of color and subtle surface geometry. However, their method requires the input to be in pairs of well-aligned before-makeup and after-makeup images and thus the practicability is limited. Guo *et al.* [10] proposed to decompose the source and reference images into face structure, skin detail, and color layers and then transfer information on each layer correspondingly. Li *et al.* [17] decomposed the image into intrinsic image layers, and used physically-based reflectance models to manipulate each layer to achieve makeup transfer. Recently, a number of makeup recommendation and synthesis systems have been developed [21, 23, 3], but their contribution is on makeup recommendation and the capability of makeup transfer is limited. As recently the style transfer problem has been successfully formulated as maximizing feature similarities in deep neural networks, Liu *et al.* [22] proposed to transfer makeup style by locally applying the style transfer technique on facial components.

In addition to makeup transfer, the problem of digitally removing makeup from portraits has also gained some attention from researchers [27, 8]. But all of them treat makeup transfer and removal as separate problems. Chang *et al.* [7] formulated the makeup transfer and removal problem as an unsupervised image domain transfer problem. They augmented the CycleGAN with a makeup reference, so that the specific makeup style of the reference image can be transferred to the non-makeup face to generate photo-realistic results. However, since they crop out the regions of eyes and mouth and train them separately as local paths, more emphasis is given to these regions. Therefore, the makeup style on other regions (such as nose, cheeks, forehead or the overall skin tone/foundation) cannot be handled properly. Very recently, Li *et al.* [18] also tackled the makeup transfer and removal problem together by incorporating "makeup loss" into the CycleGAN. Although their network structure is somewhat similar, we are the first to achieve disentanglement of makeup latent and transfer and removal on extreme and dramatic makeup styles.

Global and local discriminators. Since Goodfellow *et al.* [9] proposed the generative adversarial networks (GANs), many related works have employed discriminators in a global setting. In the domain translation problem, while a global discriminator can distinguish images from different domains, it can only capture global structures for a generator to learn. Local (patch) discriminators can compensate this by assuming independence between pixels separated by a patch diameter and modeling images as Markov random fields. Li *et al.* [16] first utilized the discriminator loss for different local patches to train a generative neural network. Such a "PatchGAN" structure was also used in [13], where a local discriminator was incorporated with an L1 loss to encourage the generator to capture local high-frequency details. In image completion [12, 19], a global discriminator was used to maintain global consistency of image structures, while a local discriminator was used to ensure consistency of the generated patches in the completed region with the image context. Azadi *et al.* [5] similarly incorporated local discriminator together with a global discriminator on the font style transfer problem.

Contrary to all previous works where only a single local discriminator is used and local patches are sampled, we incorporate multiple style discriminators specialized for different facial patches defined by facial landmarks. Therefore, our discriminators can distinguish whether the generated facial makeup style is consistent with the makeup reference, and force the generator to learn to transfer the specific makeup style from the reference.

3. LADN

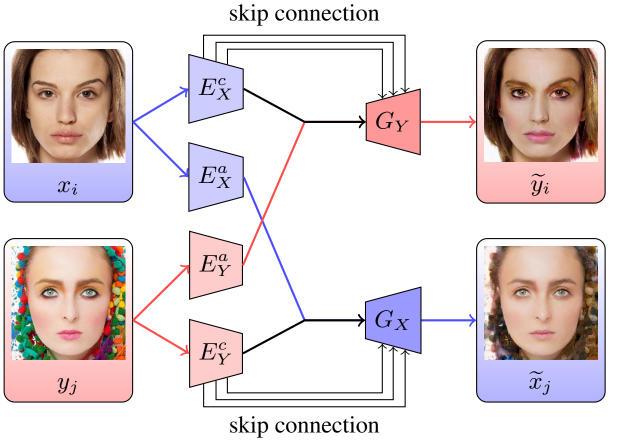


Figure 2: Generative Network Structure. The outputs of E^c and E^a are C and A, which are concatenated at the bottleneck and fed into generators. Skip connections are added between E^c and G to capture more details in generated results.

In the absence of adequate pixel-aligned before-makeup and after-makeup image datasets for our purpose, we will formulate makeup transfer and removal as an unsupervised image domain translation problem in section 3.1. We will then describe the whole network architecture and discuss our design of local style discriminators in respectively section 3.2 and section 3.3. Our asymmetric losses will be described in section 3.4 with other loss functions in the network in section 3.5.

3.1. Problem Formulation

Let image domains of before-makeup faces and after-makeup faces be $X \subset \mathbb{R}^{H\times W\times 3}$ and $Y \subset \mathbb{R}^{H\times W\times 3}$ respectively. In the unsupervised setting, we have $\{x_i\}_{i=1,\cdots,M}, x_i \in X$ to represent before-makeup examples and $\{y_j\}_{j=1,\cdots,N}, y_j \in Y$ to represent after-makeup examples, where i, j are the identities of facial images. Note that the makeup style in Y can be different for each make-up training example and there exist no before-makeup and after-makeup pairs of the same identity. The goal of the makeup transfer problem is to learn a mapping function $\Phi_Y : x_i, y_j \rightarrow \tilde{y}_i$, where $\tilde{y}_i$ receives the makeup style from y_j while preserving the identity from x_i. This can be formulated as an unsupervised cross-domain image translation problem with conditioning. The makeup removal problem can be similarly defined as $\Phi_X : y_j \rightarrow \tilde{x}_j$, an unsupervised cross-domain image translation problem from Y to X without conditioning.

3.2. Network Architecture

Recently, efforts have been put on diversifying the output of cross-domain image translation. Latest approaches leveraging disentanglement of latent variables have shown great success in similar problems [15, 4, 11, 6]. In the context of makeup transfer and removal, we want to separate the makeup style latent variable from non-makeup features

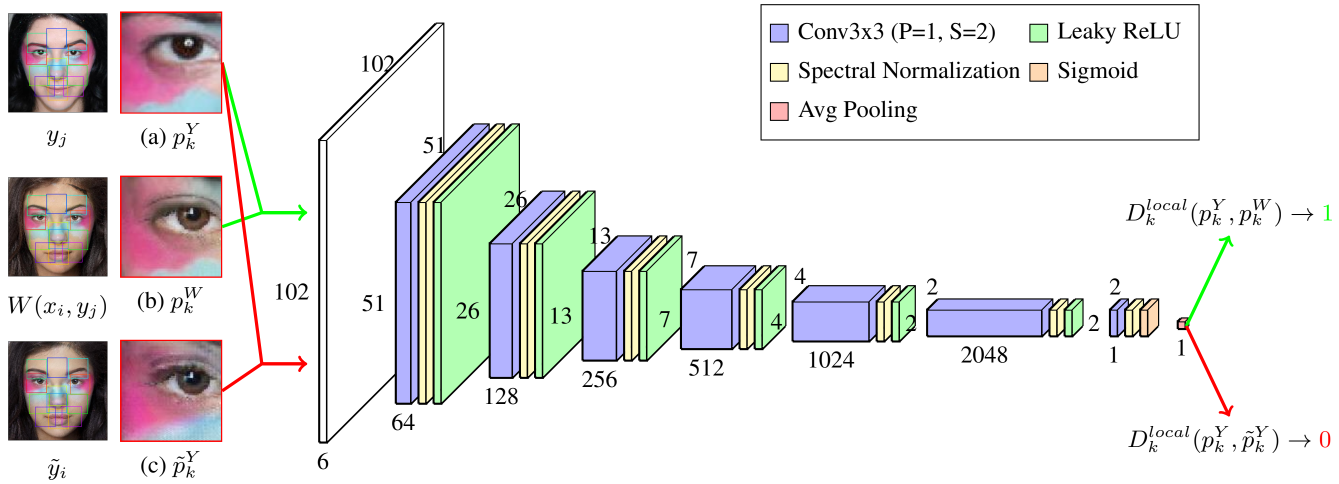


Figure 3: Local Patches and Local Discriminators. The local patches p_k^Y, p_k^W and $\tilde{p}_k^Y$ (of size $102 \times 102 \times 3$) are respectively cropped from the makeup reference, the warping result and the generated image. Pairs of p_k^Y, p_k^W are concatenated along the color channel and fed into the local discriminator as positive examples, while those of p_k^Y, $\tilde{p}_k^Y$ as negative ones. Each local discriminator is comprised of six 3×3 convolutional layers (padding=1, stride=2) with spectral normalization layers and leaky ReLU layers as shown. After the last layer of spectral normalization, the $2 \times 2 \times 1$ feature vector is passed to a sigmoid module and then averaged to produce a single scalar value which is the output indicating the probability of input pair possessing the same makeup style.

(identity, facial structure, head pose, etc.) and generate new images through recombination of these latent variables. In this process, a disentanglement framework can suppress the false correlation between makeup style and other non-makeup features. Therefore, we define attribute space A that captures the makeup style latent and content space S which includes the non-makeup features, and our network is composed of content encoders $\{E_X^c, E_Y^c\}$, style encoders $\{E_X^a, E_Y^a\}$ and generators $\{G_X, G_Y\}$.

As shown in Figure 2, by $E_X^a(x_i) = A_i$, $E_Y^a(y_j) = A_j$ and $E_X^c(x_i) = C_i$, $E_Y^c(y_i) = C_j$, we capture the attribute and content from a source image and a makeup reference, which are then fed into the generators to generate the de-makeup result $\tilde{x}_j$ and makeup transfer result $\tilde{y}_i$:

$$G_X(A_i, C_j) = \tilde{x}_j \text{ and } G_Y(A_j, C_i) = \tilde{y}_i. \quad (1)$$

The encoders and decoders are designed with a U-Net structure [24], The latent variables A, C are concatenated at the bottleneck and skip connections are used between the content encoder and generator. This structure can help retain more identity details from the source in the generated image. For the cross-domain image adaptation, we incorporate two discriminators $\{D_X, D_Y\}$ for the non-makeup domain and makeup domain, which tries to discriminate between generated images and real samples and thus helps the generators synthesize realistic outputs. And this gives the adversarial loss $L_{domain}^{adv} = L_X^{adv} + L_Y^{adv}$, where

$$\begin{aligned} L_X^{adv} &= \mathbb{E}_{x \sim P_X}[\log D_X(x)] + \mathbb{E}_{\tilde{x} \sim G_X}[\log(1 - D_X(\tilde{x}))] \\ L_Y^{adv} &= \mathbb{E}_{y \sim P_Y}[\log D_Y(y)] + \mathbb{E}_{\tilde{y} \sim G_Y}[\log(1 - D_Y(\tilde{y}))]. \end{aligned} \quad (2)$$

The discriminator D tries to discriminate between generated images and real samples, while the generator G tries to fool D and thus can learn to adapt the generated results into the target domain.

3.3. Local Style Discriminator

We propose to use multiple overlapping local discriminators to realistically transfer makeup styles which may contain high-frequency details, and this sets ourselves apart from [7] which used specialized generators and global discriminators for three key regions and thus may miss makeup details that straddle outside of those regions.

To deal with the lack of ground truth of makeup transfer y_i, inspired by [7], we generate synthetic ground truth $W(x_i, y_j)$ by warping and blending y_j onto x_i according to their facial landmarks. Although the synthetic results cannot serve as the real ground truth of the final results, they can provide guidance to the makeup transfer network on what the generated results should look like. Note that the warping results sometimes possess artifacts, which can be fixed by the network in the generated results. Based on this idea, we use local discriminators to construct style loss, which can help the generator capture the style from the makeup reference in an adversarial learning process. A typical placement of local discriminators is shown in Figure 3, where corresponding patches in the makeup reference y_j, warping reference $W(x_i, y_j)$, and generated image $\tilde{y}_i$ are marked with bounding boxes. Given the image resolution of 512×512, each local discriminator considers a local image patch of size 102×102. Note that the local discriminators are overlapping, with one discriminator trained on one key facial landmark and thus exact locations of local discriminators are not critical.

Given a set of K local discriminators $\{D_k^{local}\}_{k=1,\cdots,K}$ at each facial landmark, a local patch from the makeup reference p_k^Y (Figure 3a), the corresponding local patch from makeup warp p_k^W (Figure 3b), and that from the generated facial image $\tilde{p}_k^Y$ (Figure 3c) are cropped and fed into the local discriminator D_k^{local} in pairs. By setting the ground truth for different pairs to local discriminators, the local discriminators will learn to judge p_k^Y and p_k^W as positive pairs (of the same makeup style), and judge p_k^Y and $\tilde{p}_k^Y$ as negative pairs (of different makeup styles). Meanwhile, the goal of the generator Φ_Y is to generate the result $\tilde{y}_i$ which is of the same makeup style as the makeup reference y_j, therefore forming an adversarial learning process with the local discriminators. The loss for local discriminators is $L^{local} = \sum_k L_k^{local}$, where L_k^{local} is defined as

$$\begin{aligned} L_k^{local} =& \mathbb{E}_{x_i \sim P_X, y_j \sim P_Y}[\log D_k^{local}(p_k^Y, p_k^W)] \\ &+ \mathbb{E}_{x_i \sim P_X, y_j \sim P_Y}[\log[1 - D_k^{local}(p_k^Y, \tilde{p}_k^Y)]. \end{aligned} \tag{3}$$

The corresponding mini-max game is defined as

$$\max_{D_k^{local}} \min_{E_X^c, E_Y^a, G_Y} L^{local}. \tag{4}$$

By this setup, we take the synthetic results as guidance and encourage the local discriminators to capture makeup details from the makeup reference. Figure 3 gives the network details of local discriminators.

3.4. Asymmetric Losses

While transfer and removal of light makeup styles mainly involve re-coloring of eyeshadows and lips, extreme makeup style poses new challenges in this problem. On the one hand, extreme makeup styles contain high-frequency components, for which the network needs to differentiate from other high-frequency facial textures (e.g., eyelashes). On the other hand, in some cases of extreme makeup removal, the original facial color of the person can hardly be observed from the after-makeup image (e.g., Figure 6 (c)), which requires the network to reconstruct or hallucinate the facial skin color without makeup. To tackle these challenges, we incorporate a high-order loss L^{ho} for the makeup transfer branch to help transfer high-frequency details, and a smooth loss L^{smooth} for the de-makeup branch, based on the assumption that facial colors behind the makeups are generally smooth.

High-Order Loss: Since the warping image $W(x_i, y_j)$ preserves most texture information of the makeup style (color changes, edges) from the reference image y_j, we apply Laplacian filters to p_k^W, $\tilde{p}_k^Y$ and define high-order loss as

$$L^{ho} = \sum_k h_k ||f(p_k^W) - f(\tilde{p}_k^Y)||_1, \tag{5}$$

where h_k is the weight for local patches, and f is the Laplacian filter. We set h_k to be similar for all local patches, with slight emphases on eye regions as eye makeups can contain subtle but essential details.

Smooth Loss: Contrary to the makeup transfer result $\tilde{y}_i$, we do *not* want the de-makeup result $\tilde{x}_j$ to possess high-frequency details and instead it should be smooth in local parts. Therefore we apply a smooth loss to $\tilde{x}_j$, which is defined as

$$L^{smooth} = \sum_k s_k ||f(\tilde{p}_k^X)||_1, \tag{6}$$

where $\tilde{p}_k^X$ is a local patch from $\tilde{x}_j$, s_k is the weight for $\tilde{p}_k^X$ and f is a Laplacian filter. Different from L^{ho}, we give significantly smaller weights to eyes areas since we do not want to lose the high-frequency texture around the eyes. For cheek and nose areas, we assign larger weight and thus impose a higher degree of smoothness on these regions.

The smooth loss tries to prevent the high-frequency component from presenting in $\tilde{x}_i$, while the high-order loss tries to extract and incorporate this component into $\tilde{y}_j$. Therefore, these asymmetric losses work in tandem with each other to further improve the disentanglement of the makeup latent variable from non-makeup ones.

3.5. Other Loss Functions

Reconstruction Loss: Inspired by CycleGAN [28], we add the reconstruction losses into the network. We feed A_i, C_i into G_X to generate $\tilde{x}_i^{self}$, and A_j, C_j into G_Y to generate $\tilde{y}_j^{self}$, which should be identical to x_i and y_j respectively. This gives us self reconstruction loss. From the generated results $\tilde{x}_j$ and $\tilde{y}_i$, we again extract the attributes and contents and use them to generate $\tilde{x}_i^{cross}$ and $\tilde{y}_j^{cross}$, which should be identical to x_i and y_j. This gives us cross-cycle reconstruction loss. We use L1 loss to encourage such reconstruction consistency and define the reconstruction loss L^{recon} as

$$\begin{aligned} L^{recon} =& ||x_i - \tilde{x}_i^{self}||_1 + 8||x_i - \tilde{x}_i^{cross}||_1 + \\ & ||y_j - \tilde{y}_j^{self}||_1 + 8||y_j - \tilde{y}_j^{cross}||_1, \end{aligned} \tag{7}$$

where we use an additional scaling factor 8 for cross-cycle reconstruction loss to encourage the makeup transfer result to possess the makeup style.

KL Loss: We encourage the makeup style representation $\{A_i, A_j\}$ captured by attribute encoders $\{E_X^a, E_Y^a\}$ to be close to a prior Gaussian distribution. Therefore, we apply KL loss $L^{KL} = L_i^{KL} + L_j^{KL}$, where

$$\begin{aligned} L_i^{KL} &= \mathbb{E}[(D_{KL}(A_i||N(0,1))], \\ L_j^{KL} &= \mathbb{E}[(D_{KL}(A_j||N(0,1))], \\ \text{and } D_{KL}(p||q) &= \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx. \end{aligned} \tag{8}$$

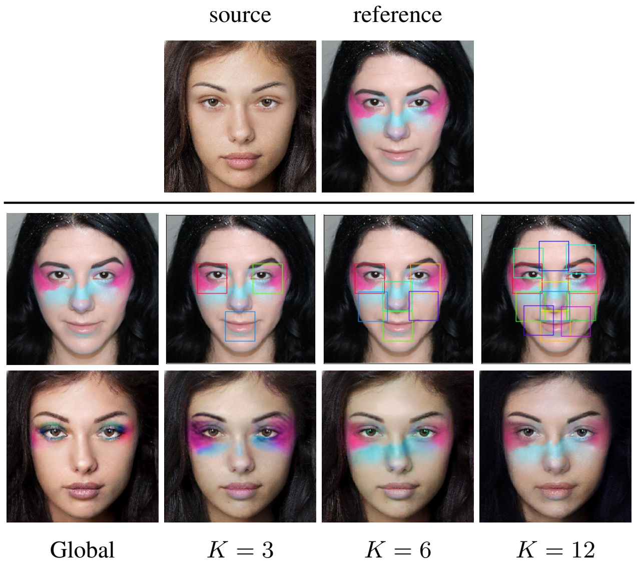


Figure 4: Results of global and local style discriminators. We evaluate the network outputs using only the global style discriminator (the first column) and 3, 6, 12 local style discriminators without the global one (the second, third and the fourth column respectively). First row: input source and reference images; second row: the placement of the local patches; third row: the makeup transfer results under different settings.

Total Loss: Our total loss is

$$\begin{aligned} L^{total} = & \lambda_{local} L^{local} + \lambda_{domain}^{adv} L_{domain}^{adv} + \lambda_{recon} L^{recon} + \\ & \lambda_{KL} L^{KL} + \lambda_{ho} L^{ho} + \lambda_{smooth} L^{smooth}, \end{aligned} \tag{9}$$

where λ_{local}, λ_{domain}^{adv}, λ_{recon}, λ_{KL}, λ_{ho}, λ_{smooth} are the weights to balance different objectives. We will provide more details of setting these weights in section 4.2.

4. Experiments

For results on conventional/light makeup transfer, removal and user studies for comparison with state-of-the-arts, please refer to the supplemental materials. In this section we focus on results on complex and dramatic makeups where no existing work had demonstrated significant results.

4.1. Data Collection

Since most datasets of face images are for recognition or identification tasks, they generally lack the labels necessary of facial makeup. There are only a few datasets on makeup that are publicly available, but most are of inadequate resolution. Some of them only contain makeup faces generated using commercial software, and thus the range of makeup styles are very limited.

As a result, we collected our own dataset, starting by collecting high-quality images of faces without occlusion from the Internet. We used facial landmark detector to filter out images without a frontal face. We then labeled a small portion of them based on the presence of makeup, from which the histogram of hue values of eyeshadow and lips regions were extracted and used to train a simple multilayer perceptron classifier. We utilized the classifier to label the remaining images and finally obtained 333 before-makeup images and 302 after-makeup images.

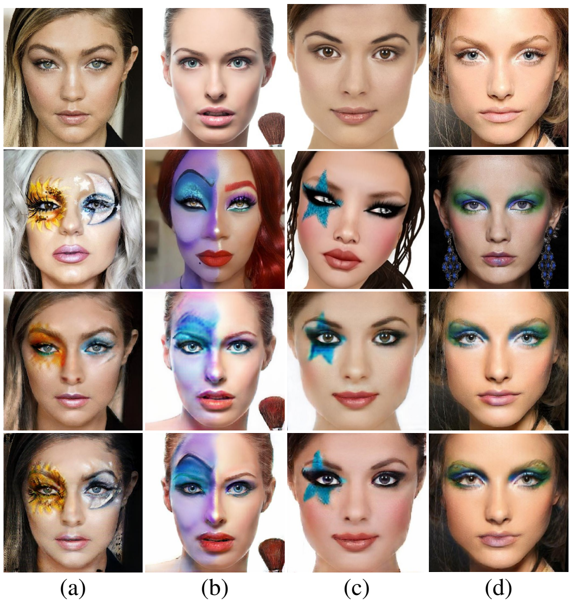


Figure 5: Makeup transfer results and ablation study on local high-order loss. First row: source images; second row: makeup references; third row: makeup transfer results from the network without local high-order loss; fourth row: makeup transfer results from the complete network.

To achieve extreme makeup transfer, we manually selected and downloaded facial images with extreme makeup by visually inspecting whether each makeup extends out of the lip and eyeshadow regions. We obtained 115 extreme makeup images with great variance on makeup color, style and region coverage, and incorporated them into the after-makeup image set.

4.2. Training Details

We incorporate $K = 12$ local discriminators into our network and set $\lambda_{local} = 2$, $\lambda_{domain}^{adv} = 1$, $\lambda_{recon} = 80$, $\lambda_{KL} = 0.01$, $\lambda_{ho} = 20$, $\lambda_{smooth} = 20$. For L^{ho}, we set h_k to 4 for areas containing eyelashes, eyelids, and 2 for areas covering nose and mouth. In L^{smooth}, s_k is set to 4 for cheek, nose areas and 0.1 for eye areas. To balance losses while number of local discriminators varies, we additionally normalize losses from the local discriminators L^{local} and losses related to local patches L^{ho} and L^{smooth} by $1/K = 1/12$. The whole network is initialized by normal initialization with $mean = 0, gain = 0.02$. We use an

Adam optimizer [14] with a learning rate of 0.001 and exponential decay rates $(\beta_1, \beta_2) = (0.5, 0.999)$. The resolution of input and output images is 512×512 and batch size is set to 1 due to the GPU memory limitation. The network is trained firstly for 700 epochs with $\lambda_{smooth} = \lambda_{ho} = 0$ to get stable with normal makeup styles, and then it is trained for 2000 epochs with $\lambda_{ho} = 20$, $\lambda_{smooth} = 20$ to boost performance on extreme and dramatic makeup styles. The input facial images are cropped and frontalized according to facial landmarks, and outputs are cropped back similarly.

4.3. Local Discriminators

To evaluate the effect of local discriminators, comparative experiments were conducted under the settings of a single global discriminator and varying numbers of local discriminators. The last row of Figure 4 shows that the network with only a single global style discriminator fails to capture the complete makeup style from the makeup reference, only adding some random color around eyes. In contrast, the network becomes focused on details of makeup style when local discriminators are incorporated (As we can tell the blue and pink texture around eyes appears for $K = 3$ compared to the Global case, which corresponds to the makeup reference). Moreover, using more local discriminators can further improve the coverage and accuracy of the transferred makeup style. As shown in Figure 4, with the expanding coverage of local discriminators ($K = 3, 6, 12$), the blue/red belt on the face gradually merges, while the texture on the nose shows stronger resemblance to the reference makeup, particularly in $K = 12$ than $K = 6$. Therefore, multiple and overlapping local discriminators are of paramount importance for our network to perform well, which makes feasible the transfer of complex makeup styles with high-frequency details covering large facial areas.

4.4. Makeup Transfer Results

As shown in Figure 5, our network can transfer the makeup styles from highly dramatic ones (Figure 5b) to those only on eyes and mouth (Figure 5d) with considerable accuracy. Although the results are not perfect as the stars in Figure 5a disappear in the transfer result, and the color on eyeshadows is a bit harsh, LADN is the first method to transfer and remove such dramatic makeup effects.

To test the effect of the high-order loss on transfer results, we ran the network with the high-order loss disabled, and the results are shown in the third row of Figure 5. Comparing the third and the fourth row, we can clearly observe that some fine details get blurred in absence of the high-order loss, and this is more severe for the makeup style with more edges (for a, b and c).

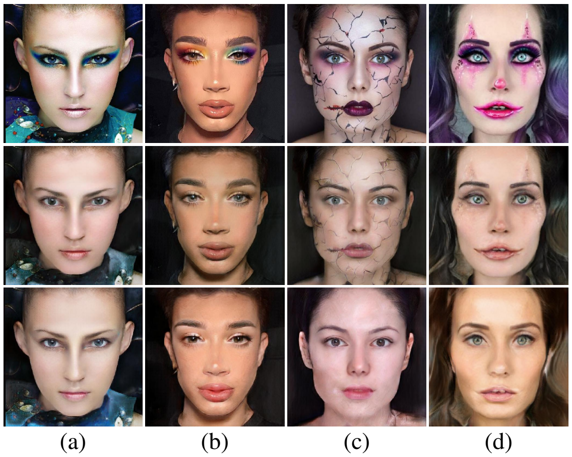


Figure 6: Makeup removal results and ablation study on smooth loss. First row: makeup reference; second row: de-makeup results from the network without smooth loss. third row: de-makeup results from the complete network for different styles, from (a)–(b) light and conventional style to (c)–(d) heavy and dramatic style.

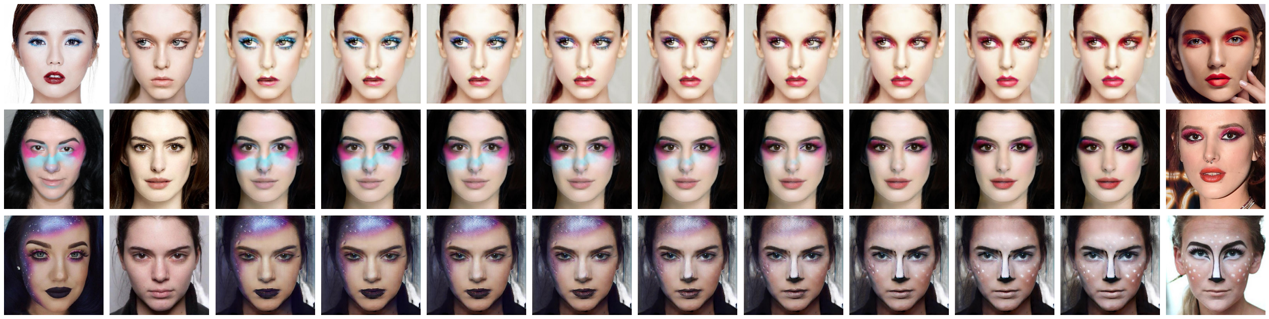
4.5. Makeup Removal Results

Figure 6 shows makeup removal results where the styles span across a spectrum from light to heavy/dramatic makeup. As discussed in section 3.4, the makeup removal problem is ill-posed in the sense that there can be multiple possible faces behind the same makeup style. This can be reflected by the makeup reference in Figure 6c, where the makeup style covers almost the whole face and the network has no clue about the face color behind the makeup from the given image. Using the asymmetric losses, our network succeeds in distinguishing makeup style from facial texture and removing it. Then the generator hallucinates to give the identity a reasonable skin color.

We also demonstrate the efficacy of smooth loss on extreme makeup removal by the ablation study in Figure 6. Similar to the high-order loss, smooth loss demonstrates its significance especially when the makeup involves edges striding over large areas out of the mouth and eyes areas (for c and d). Meanwhile, the network can also generate satisfactory light makeup removal results (Figure 6a and 6b). The applied eyeshadow and the lipstick are removed, recovering the normal face color without significant changes to other no-makeup areas.

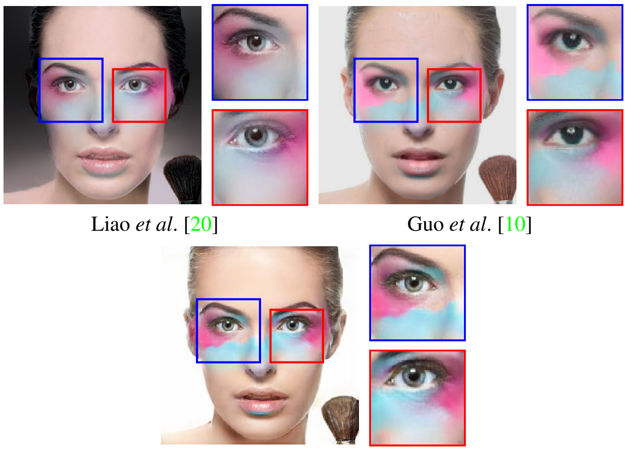
4.6. Qualitative Comparison

To our best knowledge, we are the first to achieve transfer and removal of dramatic makeup styles requiring no extra inputs. Figure 8 shows a qualitative comparison on extreme makeup transfer between our results and those from [10] and [20]. For [20], we show the result after refinement step, because otherwise the identity of the source will be lost. But as shown the refinement also blurs the makeup details. Similarly, the result from [10] depicts artifacts such as dis-



Reference 1 Source Interpolated results Reference 2

Figure 7: Interpolated makeup styles.



Liao *et al*. [20] Guo *et al*. [10]

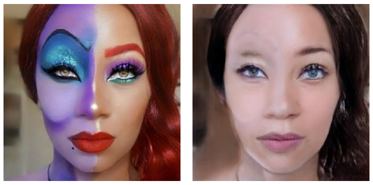
Ours

Figure 8: Comparison on extreme makeup transfer.

continuity along the boundary of skin area and color fading of the makeup. Note the method in [10] is based on conventional methods which requires extreme accuracy on face geometry alignments between two faces. In contrast, our method only requires roughly correct landmarks in order to define the location of local discriminators.

4.7. Interpolated Makeup Styles

With LADN, the attribute space is disentangled well from the content space, and we can therefore easily obtain intermediate makeup styles by interpolating two attribute vectors. Given attribute A_1 and A_2 respectively extracted from makeup reference 1 and 2, we compute $\alpha A_1 + (1-\alpha) A_2$ $(\alpha \in [0, 1])$, and feed the resulting composite attribute into the generator to yield smooth transition between two reference makeup styles. Figure 7 shows the interpolated results from left to right which depicts a smooth and natural transition: gradual increase on the lip color and red eyeshadow as well as a gradual decrease on the bluish-pink extreme makeup, without affecting the facial identity and facial structure of the source. Such interpolation capability enables our LADN network to not only control the amount or heaviness of the generated makeup but also to mix two makeup styles to generate a new style as shown, thus significantly broadening the range of styles through this simple mix-and-match feature provided by LADN.



Makeup reference De-makeup result

Figure 9: Limitation. One de-makeup example in the dataset.

5. Limitations and Conclusion

One limitation of our network is that it struggles to remove extreme makeup styles where colors are highly consistent in local regions but vary sharply across local patches. Figure 9 shows such style which divides the face into two halves, with each half coherent within itself (purple on left and orange on right). With very few high-frequency details present, the smooth loss is unable to take effect. As a result, the de-makeup network produces a plausible face behind the given makeup for *each* half of the face. Moreover, our smooth loss is designed to encourage local smooth color transition in the de-makeup result, which is different from existing de-makeup methods which aims to recover facial imperfections before the light cosmetic makeup. Consequently, our smooth loss removes the mole as well, which in hindsight may be part of the dramatic makeup as well.

In conclusion, we propose the Local Adversarial Disentangling Network (LADN) by incorporating local style discriminators, disentangling representation and asymmetric loss functions into a cross-domain image translation network. We apply LADN to makeup transfer and removal, which demonstrates its power in transferring extreme makeup styles with high-frequency color changes and details covering large facial areas, which cannot be handled by previous work. Our network also achieves state-of-the-art performance in transferring and removing light/typical makeup styles. We believe this framework can also be applied to applications beyond makeup transfer and removal, which is a fruitful future research direction to explore.

References

[1] Meitu - beauty themed photo & video apps. http://global.meitu.com/en/products. Accessed: 2018-10-02. 2

[2] Taaz virtual makeover & hairstyles. http://www.taaz.com/. Accessed: 2018-10-02. 2

[3] Taleb Alashkar, Songyao Jiang, Shuyang Wang, and Yun Fu. Examples-rules guided deep neural network for makeup recommendation. In *AAAI Conference on Artificial Intelligence*, 2017. 2

[4] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 195–204, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 3

[5] Samaneh Azadi, Matthew Fisher, Vladimir Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content gan for few-shot font style transfer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[6] Jinming Cao, Oren Katzir, Peng Jiang, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. Dida: Disentangled synthesis for domain adaptation. *arXiv preprint arXiv:1805.08019*, 2018. 3

[7] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4

[8] Ying-Cong Chen, Xiaoyong Shen, and Jiaya Jia. Makeup-go: Blind reversion of portrait edit. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4511–4519, Oct 2017. 2, 3

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 2, 3

[10] Dong Guo and Terence Sim. Digital face makeup by example. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 73–79, June 2009. 2, 7, 8

[11] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3

[12] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 36(4):107:1–107:14, 2017. 2, 3

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, July 2017. 2, 3

[14] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 7

[15] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *2018 European Conference on Computer Vision (ECCV)*, September 2018. 2, 3

[16] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *2016 European Conference on Computer Vision (ECCV)*, 2016. 2, 3

[17] Chen Li, Kun Zhou, and Stephen Lin. Simulating makeup through physics-based manipulation of intrinsic image layers. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4621–4629, June 2015. 2

[18] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 645–653, 2018. 3

[19] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5892–5900, July 2017. 2, 3

[20] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *ACM Trans. Graph.*, 36(4):120:1–120:15, July 2017. 7, 8

[21] Luoqi Liu, Hui Xu, Junliang Xing, Si Liu, Xi Zhou, and Shuicheng Yan. "wow! you are so beautiful today!". In *Proceedings of the 2013 ACM International Conference on Multimedia*, MM '13, pages 3–12, New York, NY, USA, 2013. ACM. 2

[22] Si Liu, Xinyu Ou, Ruihe Qian, Wei Wang, and Xiaochun Cao. Makeup like a superstar: Deep localized makeup transfer network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2568–2575. AAAI Press, 2016. 2

[23] Tam V. Nguyen and Luoqi Liu. Smart mirror: Intelligent makeup recommendation and synthesis. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 1253–1254, New York, NY, USA, 2017. ACM. 2

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 4

[25] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[26] Wai-Shun Tong, Chi-Keung Tang, Michael S Brown, and Ying-Qing Xu. Example-based cosmetic transfer. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 211–218, Oct 2007. 2

[27] Shuyang Wang and Yun Fu. Face behind makeup. In *AAAI Conference on Artificial Intelligence*, AAAI'16, pages 58–64. AAAI Press, 2016. 2, 3

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, Oct 2017. 2, 5