Semi-supervised Eye Makeup Transfer by Swapping Learned Representation

Feida Zhu[1,2], Hongji Cao[2], Zunlei Feng[3], Yongqiang Zhang[2], Wenbin Luo[2], Hucheng Zhou[2], Mingli Song[3], and Kai-Kuang Ma[1]

[1]Nanyang Technological University, [2]Alibaba Group, [3]Zhejiang University
{feida.zhu, EKKMA}@ntu.edu.sg

Abstract

This paper introduces an autoencoder structure to transfer the eye makeup from an arbitrary reference image to a source image realistically and faithfully using both synthetic paired data and unpaired data in a semi-supervised way. Different from the image domain transfer problem, our framework only needs one domain entity and follows an "encoding-swap-decoding" process. Makeup transfer is achieved by decoding the base representation from a source image and makeup representation from a reference image. Moreover, our method allows users to control the makeup degree by tuning makeup weight. To the best of our knowledge, there is no public large makeup dataset to evaluate data-driven approaches. We have collected a dataset of non-makeup images and with-makeup images of various eye makeup styles. Experiments demonstrate the effectiveness of our method with the state-of-the-art methods both qualitatively and quantitatively.

1. Introduction

Makeup is to enhance a unique theme with special cosmetics, such as foundation, eye brush, eye shadow and etc., to make the person more attractive. There have been arising needs for realistic digital makeup from the makeup industry. Consider this scenario: when a customer shops on-line, she browses the website and is attracted by a model's makeup. Naturally, the customer would like to try out this makeup on her face before purchase. Automatic digital makeup transfer from the model to the customer's face is favorable. Some APPs like ModiFace and Taaz can digitally add cosmetic elements to a face in a photo. However, the styles are limited to a pre-determined cosmetic set. Among different makeup cosmetics, eye makeup transfer is especially challenging since eye makeup not only involves color and texture changes such as eyeshadow, but also contains high frequency detail changes such as eyelashes.

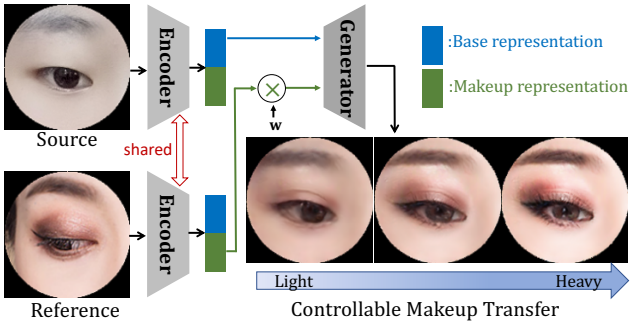

Figure 1: At inference stage, the source image can present eye makeup styles of the reference image with controllable makeup degree. The weight **w** is gradually increased to demonstrate that the makeup degree can be controlled conveniently

This paper introduces an autoencoder architecture to digitally add eye makeup to a source image where the eye makeup style is consistent with the reference image (Figure 1). In physical makeup process, we can assume that a makeup layer is added to the base layer to beautify the skins while preserving the source structure. Based on this assumption, the proposed autoencoder is trained to extract base representation and makeup representation on input images. The makeup transfer result is obtained by decoding the base representation from a source image and makeup representation from a reference image. Compared to the state-of-the-art makeup transfer method [5] where the transfer is accomplished within a black-box generator, the makeup degree is controllable in our method by tuning the makeup weight thanks to the learned disentangled representation.

One challenge for data-driven approaches is that it's difficult to obtain triplets training data: the source, the reference with makeup and the ground truth output (which preserves the identity of the source and present the makeup of the reference). Due to the lack of such data, we formulate the makeup transfer problem under unpaired setting like

unsupervised image-to-image translation [37]. GAN-based makeup transfer method for unpaired data [5] can constrain the generated output to lie on the manifold of real examples, but it can not guarantee the makeup is transferred from the reference image faithfully. To encourage faithful makeup transfer, we synthesize "ground truth" makeup transferred results by image warping based on Delaunay triangulation [22] to facilitate supervised learning. The semi-supervised training scheme turns out to be effective in encouraging the synthetic paired data and unpaired data to reinforce each other to produce both faithful and realistic makeup transfer results.

The whole architecture of our method is shown in Figure 2. It consists of two "encoding-swap-decoding" processes which use the same autoencoder. Let X and Y denote the non-makeup and with-makeup images. The synthetic paired data $(X, X^*), (Y, Y^*)$ only go through the first encoder-swap-decoder process, where pixel-wise constraints are introduced to ensure the makeup being transferred faithfully. It is notable that the synthetic images X^*, Y^* are far from real examples and pixel-wise losses tend to produce smoothed results. In comparison, the original unpaired data X, Y go through the two encoding-swap-decoding processes consecutively. Cycle consistent loss and adversarial loss are introduced to encourage the network to generate realistic results.

To the best of our knowledge, there is no public large makeup dataset of high resolution to evaluate data-driven approaches. We have collected a dataset of 1000 images without makeup and 1000 images with eye makeup.

The key contributions of this paper are: 1) The proposed autoencoder framework combines the unpair samples and synthetic paired samples to extract disentangled base representation and makeup representation, which can generate faithful and realistic makeup transfer results. 2) With the disentangled makeup representation, our framework allows users to control the makeup degree by simply tuning the makeup weight. 3) We collect the first makeup dataset, which contains non-makeup images and with-makeup images of various eye makeup styles. Our code, data and results will be made public.

2. Related Work

Facial Makeup Transfer: Makeup transfer requires to transfer the makeup component of the reference image (such as eyeshadow, eyelashes) to the corresponding regions of source face precisely and realistically. Several previous works [33, 14, 24] attempted to address the challenges by transferring the appearance statistics in each decomposed layer separately with hand-crafted features. Liu *et al.* [27] applied the optimized-based neural style transfer model [12] with local constraints to transfer different cosmetics in different manners. A major disadvantage is that it needs to parse the face into different semantic regions first, which is prone to error. The state-of-the-art deep method proposed by Chang *et al.* [5] extends the CycleGAN [37] to asymmetric networks to enable transferring specific style from reference image. However, the transfer process is completed within a black-box network without any control of how much makeup being transferred. In contrast, we propose an autoencoder to extract well-disentangled base and makeup representation with semi-supervised training scheme. Faithful and realistic makeup transfer is achieved by decoding the base representation from a source image and makeup representation from an arbitrary reference makeup image. In addition, the makeup degree can be controlled conveniently by tuning the makeup representation weight.

Disentangled Representation: Disentangling aims at learning disentangled representations from data. Existing works in learning disentangled representations can be categorized into two groups, unsupervised methods and (semi-)supervised methods. Most of the existing unsupervised methods [4, 6, 11, 9] are based on β-VAE [16], which strength the prior isotropic Gaussian distribution of latent code to achieve disentangling. Base on VAE, FactorVAE [19] adds the total correlation into β-VAE, which can get better disentangled representation. Other unsupervised disentangling methods are based on InfoGAN [7], which is an information-theoretic extension to the Generative Adversarial Network(GAN) that is able to learn disentangled representations in a completely unsupervised manner. Recently, Locatello *et al.* [28] challenge common assumptions in the unsupervised learning of disentangled representations. Their work theoretically shows that the unsupervised learning of disentangled representations is fundamentally impossible without inductive biases on both the models and the data. Existing (semi-)supervised methods can be roughly divided into three kinds. Some semi-supervised methods [3, 30] import annotation information into β-VAE to achieve controllable disentangling. Another direction [2, 20, 31, 34] is to utilize annotated data to supervise the input-to-attribute mapping explicitly. The third kind methods [8, 10, 13] consider to combine paired images to achieve disentangled representation implicitly in an end-to-end training manner. However, the disentangling performance of above (semi-)supervised methods strongly depends on the annotated samples.

Image Domain Adaption: A series of works including CycleGAN [37], DualGAN [36], UNIT [26] were proposed for general image-to-image translation problem under the unsupervised setting. However, these methods are limited to deterministic one-to-one mapping. They can not replicate a specific makeup style, which makes them not applicable in this makeup transfer scenario. Most recently, MUNIT [18] and several concurrent works [1, 23] adopt dis-

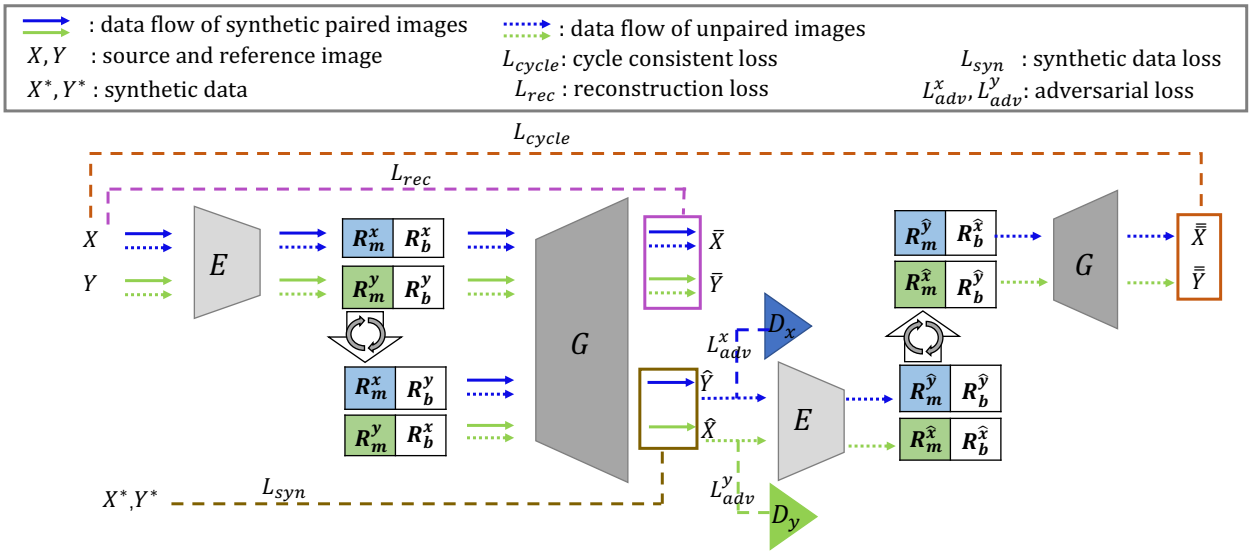

Figure 2: The architecture of our method. Both the synthetic paired data and unpaired data go through the first encoder E and generator G while only unpaired data go through the second E and G. Synthetic data loss L_{syn} and reconstruction loss L_{re} are designed for synthetic paired data. Cycle consistent loss L_{cycle} and adversarial loss L_{adv} are designed for unpaired data.

entangled representations to generate multiple results with various styles. However, MUNIT only generates texture changes and fails to preserve the source image structure in our makeup transfer application. Different from previous works which require coupled generators to translate images between two domains, our method operates in one domain with one autoencoder.

3. Method

Due to the lack of well-aligned non-makeup and with-makeup image pairs, we formulate our problem at unpaired setting. Let X and Y denote the non-makeup and with-makeup image domains. Different from previous GAN-based image translation model [18], we apply an autoencoder structure to learn disentangled base representation and makeup representations. The considerations are twofold. The first is motivated by the physical makeup process, which can be regarded as a makeup layer being applied to the base structure layer. The second is to propose a more interpretable and controllable method rather than a black-box network. Moreover, semi-supervised learning scheme is used to facilitate the training of the autoencoder. The supervised training for synthetic paired data and the unsupervised training for unpaired data can reinforce each other to produce faithful and realistic makeup transfer result, as will be illustrated in details in Section 3.2.

To sum up, our method learns to decompose an input image into base representation R_b, representing the underlying base structure information, and makeup representation R_m, representing the makeup rendering information. The makeup-transferred result is generated by decoding the base representation of a source image and the makeup representation of a reference image.

3.1. Network Architecture

Figure 2 illustrates the overall architecture, which includes encoder E, generator G and discriminator D_x, D_y. The encoder E maps the image into makeup representation and base representation $[R_m, R_b] = E(x)$. We expect the representation to have the following two properties. (1) They should contain the information of input as much as possible and can be inverted back. (2) R_m and R_b are well disentangled. To achieve the first property, reconstruction loss L_{re} is introduced to penalize the difference between reconstructed image $\overline{x}, \overline{y}$ and origin image x, y. To achieve the second property, the makeup representation are swapped to generate makeup-transferred image $\hat{x} = G(R^y_m, R^x_b)$ and makeup-remove image $\hat{y} = G(R^x_m, R^y_b)$. For synthetic paired data, $\hat{x}, \hat{y}$ should be close to the synthetic "groundtruth" images x^*, y^*. For unpaired data, $\hat{x}, \hat{y}$ will go through the second "encoding-swap-decoding" process. Adversarial losses L_{adv} are designed to encourage the generated results indistinguishable from the real images. We also inherit the cycle consistency loss L_{cycle} from CycleGAN [37] to preserve the image identity after dual representation swap. That is, $\overline{\overline{x}} = G(R^{\hat{y}}_m, R^{\hat{x}}_b), \overline{\overline{y}} = G(R^{\hat{x}}_m, R^{\hat{y}}_b)$ should be identical to x, y respectively.

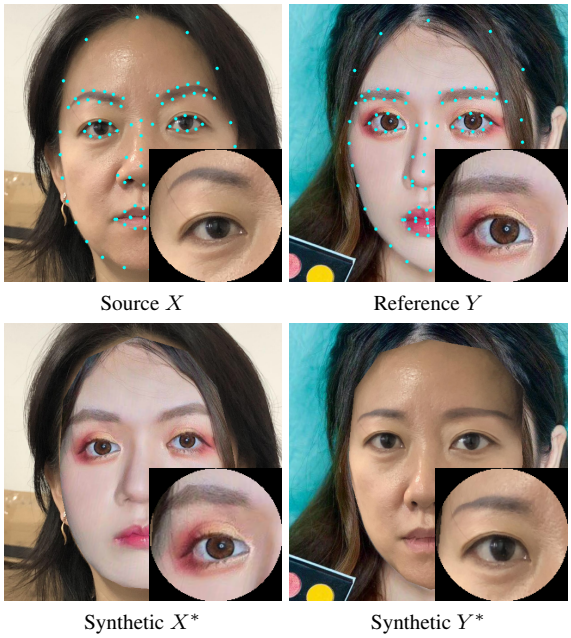

Figure 3: We synthesize warped image X^*, Y^* to facilitate supervised training with paired data, (X, X^*) and (Y, Y^*). Although the person's identity is lost during warping (*e.g.*, the eyelid is different obviously), paired images can offer strong pixel-wise constraints on autoencoder.

3.2. Training Pipeline

For Synthetic Paired Data: To drive the generator to generate accurate makeup style close to the reference image, we synthesize ground-truth makeup transferred image X^* by warping the reference image Y to match the facial landmarks of source image X based on Delaunay triangulation [22]. Y^* is produced in a similar way which denotes the makeup removal result. An example is shown in Figure 3 where we can see that the identity of X, Y is lost in the warped image X^*, Y^*. For example, the eyelid is different obviously. Although the synthetic paired data (X, X^*) and (Y, Y^*) are fake, they can offer strong pixel-wise constraints on autoencoder, which is achieved by minimizing the following synthetic data loss:

$$\begin{aligned} L_{syn}(E,G) &= \mathbb{E}_{x\sim\mathcal{P}_X, y\sim\mathcal{P}_Y}[\|G(R_m^y, R_b^x) - x^*\|_1] \\ &= \mathbb{E}_{x\sim\mathcal{P}_X, y\sim\mathcal{P}_Y}[\|G(R_m^x, R_b^y) - y^*\|_1], \end{aligned} \quad (1)$$

where $[R_m^x, R_b^x] = E(x), [R_m^y, R_b^y] = E(y)$.

The autoencoder is expected to reconstruct the original input image based on the encoded representation. That is, $\overline{x}, \overline{y}$ should be identical to x, y correspondingly. This gives us self reconstruction loss. The reconstruction loss $L_{rec}(E, G)$ is defined as:

$$\begin{aligned} L_{rec}(E,G) &= \mathbb{E}_{x\sim\mathcal{P}_X, y\sim\mathcal{P}_Y}[\|G(R_m^x, R_b^x) - x\|_1] \\ &+ \mathbb{E}_{x\sim\mathcal{P}_X, y\sim\mathcal{P}_Y}[\|G(R_m^y, R_b^y) - y\|_1]. \end{aligned} \quad (2)$$

Algorithm 1 The Training Algorithm

Input 1: synthetic paired data: $(x_i, x_i^*), (y_i, y_i^*), i = 1, .., M$
Input 2: unpaired data: $x_j, y_j, j = 1, ..., N$
Model: Encoder E, Generator G, Discriminator D_x, D_y
Output: Encoder E, Generator G,

Initialize E, G, D_x, D_y
for $t = 1, 2...T$ iteration **do**
  Random sample paired data $(x, x^*), (y, y^*)$
  Encode x, y into representations: $R_m^x, R_b^x, R_m^y, R_b^y$
  Reconstruct $\overline{x}, \overline{y}$
  Swap makeup representation and generate $\hat{x}, \hat{y}$
  Update E, G by optimizing $\min_{E,G} L_s(E, G)$
  Random sample unpaired data x, y
  Encode x, y into representation: $R_m^x, R_b^x, R_m^y, R_b^y$
  Reconstruct $\overline{x}, \overline{y}$
  Swap makeup representation and generate $\hat{x}, \hat{y}$
  Apply discriminator D_x, D_y on $\hat{x}, \hat{y}$
  Encode $\hat{x}, \hat{y}$ into representations: $R_m^{\hat{x}}, R_b^{\hat{x}}, R_m^{\hat{y}}, R_b^{\hat{y}}$
  Swap back base representation and generate $\overline{\overline{x}}, \overline{\overline{y}}$
  Update E, G, D_x, D_y by optimizing $\min_{E,G} \max_{D_x, D_y} L_u(E, G, D_x, D_y)$
end for

Due to the nature of non-makeup image, we encourage its makeup representation to be sparse, and add a regularization term $L_{regu}(E) = \|R_m^x\|_1$. In summary, we train the encoder and generator to optimize the following objective function on total supervised loss L_s for synthetic paired data $(X, X^*), (Y, Y^*)$:

$$\min_{E,G} L_s(E,G) = \lambda_{rec} L_{rec} + \lambda_{syn} L_{syn} + \lambda_{regu} L_{regu}, \quad (3)$$

where λ_{rec}, λ_{syn} and λ_{regu} balance the multiple objectives.

For Unpaired Data: Unlike the paired data that only go through the first "encoding-swap-decoding" process, the unpaired data X, Y go through the process twice consecutively. The total unsupervised loss for unpaired data L_u consists of three types: reconstruction loss L_{rec} as described in Equation 2, adversarial loss L_{adv} and cycle consistent loss L_{cycle}.

The adversarial loss constrains the output of generator $\hat{x}, \hat{y}$ to lie on the manifold of Y, X respectively. Specifically, let us take $\hat{x}$ as an example. Discriminator D_y is to distinguish the generated result $\hat{x}$ from the real ones in Y. The generator aims to generate indistinguishable results conditioned on the base representation of source image R_b^x and makeup representation of reference image R_m^y. The adversarial loss is defined as:

$$\begin{aligned} L_{adv}^x(E,G,D_x) &= \mathbb{E}_{x\sim\mathcal{P}_X}[log(D_x(x))] \\ &+ \mathbb{E}_{x\sim\mathcal{P}_X, y\sim\mathcal{P}_Y}[log(1 - D_x G(R_m^x, R_b^y))], \end{aligned} \quad (4)$$

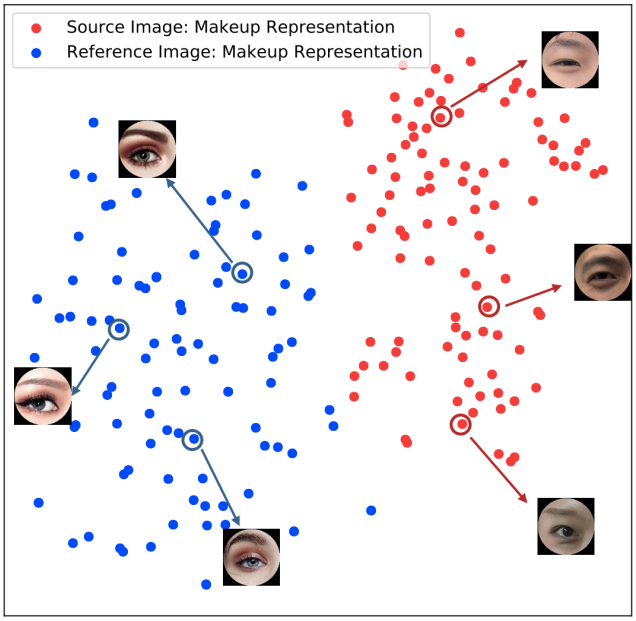

Figure 4: t-SNE visualization of makeup representation of all test images where they are placed exactly at their embedded location.

$$L^y_{adv}(E,G,D_y) = \mathbb{E}_{y\sim\mathcal{P}_Y}[log(D_y(y))] + \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}[log(1-D_xG(R^y_m,R^x_b))]. \quad (5)$$

The cycle consistent loss is to preserve the identity of x, y. That is, we should get back the original image x, y exactly if we swap back the makeup representation in the second "encoding-swap-decoding" process. Specifically, $\hat{x}, \hat{y}$ are fed to the same encoder E again and we obtain their representation $R^{\hat{x}}_m, R^{\hat{x}}_b, R^{\hat{y}}_m, R^{\hat{y}}_b$ as shown in Figure 2. We swap back the makeup representation and decode them to $\overline{\overline{x}} = G(R^{\hat{y}}_m, R^{\hat{x}}_b), \overline{\overline{y}} = G(R^{\hat{x}}_m, R^{\hat{y}}_b)$, which should be identical to x, y. The cycle consistent loss is defined as:

$$L_{cycle}(E,G) = \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}[\|G(R^{\hat{y}}_m, R^{\hat{x}}_b) - x\|_1] + \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}[\|G(R^{\hat{x}}_m, R^{\hat{y}}_b) - y\|_1], \quad (6)$$

where $[R^{\hat{x}}_m, R^{\hat{x}}_b] = E(\hat{x}), [R^{\hat{y}}_m, R^{\hat{y}}_b] = E(\hat{y})$.

In summary, we train the encoder, generator and discriminator to optimize the following objective function on total unsupervised loss L_u for unpaired data X, Y:

$$\min_{E,G}\max_{D_x,D_y} L_u(E,G,D_x,D_y) = \lambda_{rec}L_{rec} + \lambda_{cycle}L_{cycle} + \lambda_{adv}(L^x_{adv} + L^y_{adv}), \quad (7)$$

where $\lambda_{rec}, \lambda_{cycle}$ and λ_{adv} balance the multiple objectives.

Training Algorithm: The complete training algorithm is summarized in Algorithm 1. Within each iteration during training, we optimize the network alternatively using synthetic paired data and unpaired data. Training supervised loss is more stable than unsupervised loss. In each iteration, optimizing supervised loss first can provide a better starting point of encoder E and generator G before optimizing unsupervised loss so that the discriminator D can be trained more efficiently. As a result, the whole model converges faster. After training, t-SNE [29] is utilized to compute a two dimensional embedding of makeup representation R_m. The embedding arranges eyes with similar makeup representation close to each other in a plane. Figure 4 shows the resulting spatial layout, from which we can see the well-disentangled makeup representation clearly divides the images into two groups.

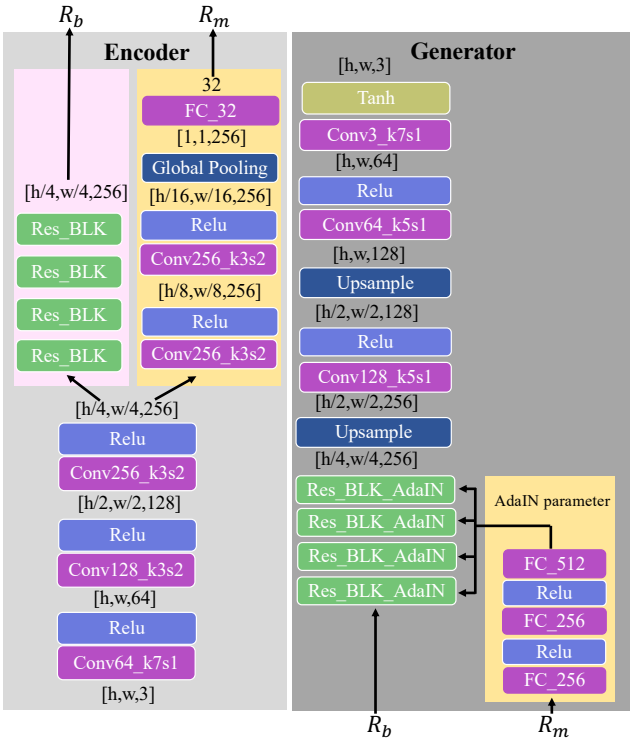

Figure 5: Architecture of encoder and generator network. The convolutional channel number, kernel size and stride are indicated for each convolutional layer. We have designed four basic residual blocks [15] in the encoder and four residual blocks with adaptive instance normalization [17] in the generator to process the disentangled representation.

4. Experiment

4.1. Implementation Details

Data Collection: Existing face datasets collected for recognition tasks generally lack clear makeup. Besides, the faces with makeup and without makeup are mixed together. Thus, we collected a new dataset for our semi-supervised learning problem. We first used facial landmark detector to collect high-quality frontal face images from the Internet. We then

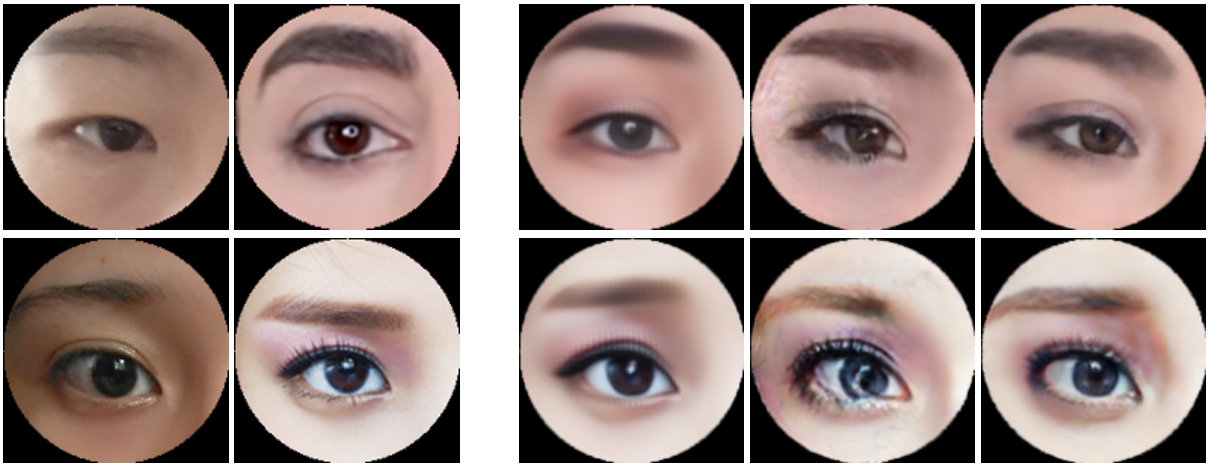

Figure 6: Comparison of visual results trained with different loss settings. The semi-supervised scheme which utilizes supervised loss and unsupervised loss together achieves more natural results than those obtained using one type of loss only.

manually selected with-makeup and no-makeup images by visually inspecting whether the eye region contains obvious makeups. Ambiguous images were discarded. In this way, we obtained 1000 with-makeup images and 1000 non-makeup images. Our dataset contain a wide variety of identities and eye makeup styles.

Model Details: The network architecture is illustrated in Figure 5. The encoder applies several stride-2 convolutional layers to downsample the input image and is further divided into two branches. Base representation maintains the spatial resolution to avoid shape information loss while makeup representation is downsampled to a 32-dimensional vector to contain style information only. Generator processes the base representation by several residual blocks equipped with adaptive instance normalization [17] to reconstruct the final image output conditioned on makeup representation. The training pipeline has been illustrated in Algorithm 1. We employ the discriminator proposed by [35] to guide the generator, which consists of four stride-2 convolutional layers and four leaky ReLU layers.

Training Setting: The reconstruction loss and cycle consistent loss are more crucial since the model should retrieve images accurately. We set $\lambda_{rec} = \lambda_{cycle} = 1$, $\lambda_{adv} = \lambda_{syn} = 0.5$ and $\lambda_{regu} = 0.1$. We tested different parameter settings systematically and chose the hyperparameters according to the qualitative results. The results maintained high quality even if the loss weights changed within a certain range. 900 with-makeup and 900 no-makeup images are randomly selected for training while the remaining images are used for testing. Random horizontal flip and random rotation from -30 degree to 30 degree are applied as data augmentation. The network is trained for 400 epochs, which take around two days using Tesla P100 GPU. At the inference stage, the makeup transfer is achieved within 0.2s.

4.2. Effectiveness of Semi-supervised Training Scheme

Our network is trained with supervised loss L_s and unsupervised loss L_u iteratively as described before. We conduct experiments using three different loss settings, including L_s alone, L_u alone and L_s, L_u together to analyze the effect of each loss term. The training hyperparameters are kept the same. We provide two examples of visual comparison in Figure 6. When we train the network in supervised manner only, the results are faithful to the synthetic data X^*. However, the results are overly-smoothed and have poor perceptual quality. A similar finding of the limitation of pixel-wise mean square error loss has also been reported in [21]. When we train the network in an unsupervised manner only, the faithfulness can't be guaranteed. As Figure 6 shows, the eyelashes are not transferred accurately. In contrast, the combined semi-supervised scheme has produced better results than those obtained using one type of loss only.

4.3. Comparison with State-of-the-art

In Figure 7, we compare with three state-of-the-art methods, MUNIT [18], DeepImageAnalogy [25] and PairedCycleGAN [5]. MUNIT [18] is the state-of-the-art unsupervised image-to-image translation method which can learn multimodal mappings between two visual domains. However, it fails to preserve the structure in our makeup adaption problem without the guidance of supervised constraints. DeepImageAnalogy [25] transfers style in a structure preserving manner by finding dense semantic correspondences in feature space of pre-trained VGG19 [32]. However,

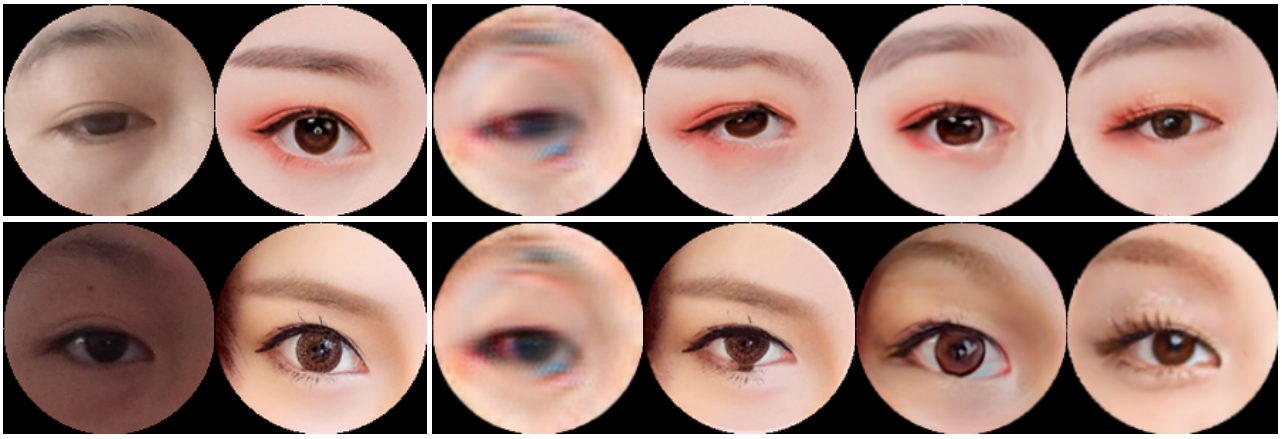

Figure 7: Makeup transfer results. We compare our method with state-of-the-art general style transfer methods [18, 25] and makeup transfer method [5].

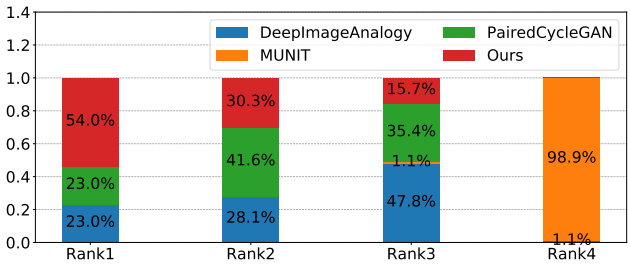

Figure 8: Ranking statistics of each method. People prefer our results over those of other three methods.

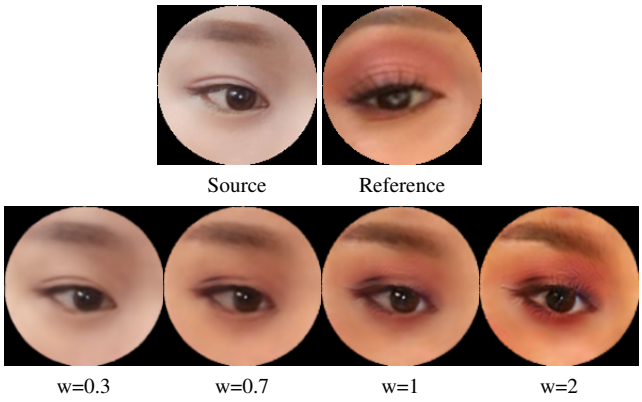

Figure 9: Learned disentangled representations allow users to control the makeup degree by tuning makeup representation weight easily.

VGG19 is originally trained for object detection problem where the data statistics vary greatly from our eye makeup dataset. This inherent contradiction decreases the semantic correspondence accuracy. Comparing the first column and fourth column of Figure 7, the eyebrow of the source image is not consistent with that of the makeup transferred result. PairedCycleGAN [5] is the state-of-the-art fast style transfer method aimed at makeup transfer. Their method's main limitation is that the transfer process is not controllable at inference stage. In contrast, our method encodes the makeup style into a 32-dimensional makeup representation. The makeup degree can be adjusted by tuning the makeup representation weight. Besides, some unwanted statistics of the reference image are transferred to the source image by their method. For example, the eye size of the reference image is transferred to the source image mistakenly, as shown in Figure 7.

Quantitative Comparison:

We conduct a user study to assess the makeup transfer quality of different approaches. The study shows the source image and reference makeup image at the top. Four results generated by MUNIT [18], DeepImageAnalogy [25], PairedCycleGAN [5] and our method are presented to the participants in random order at the bottom. The participants are asked to rank them in terms of preserving the source image identity and matching the reference image makeup style. We use 50 questions and collect 9 or more responses to each question. As shown in Figure 8, our method is chosen as the best 54.0% of the times. The average rank of our method is 1.62, which outperforms all other methods.

4.4. Makeup Control

One advantage of our method is that the network has learned interpretable makeup representation. Let us introduce the makeup representation weight w and generate makeup transferred result by $G(w * R_m^y, R_b^x)$. The weight is gradually increased to demonstrate that the makeup de-

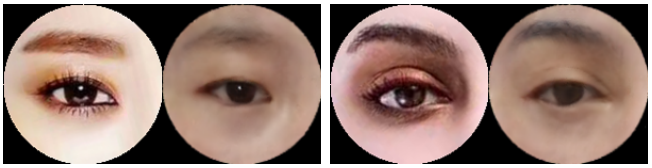

Figure 10: Makeup removal results.

gree can be controlled conveniently. Generated results with different weights are shown in Figure 9. The eyebrow, eyeshadow and eyelash become more vivid with larger weights. Controllable makeup representation makes our method have boarder range of applications than current state-of-the-art makeup transfer method [5] which can only replicate a specific style from the reference image. The reason behind the success of applying our method to other problems lies in that our well-designed encoder-generator architecture can extract disentangled representation to encode the base structure and texture information. Therefore, the style transfer process is more controllable and explicable.

4.5. Makeup Removal

Recover of the original face behind makeup is an ill-posed problem since the makeup process may conceal the original appearance. For example, the foundation makeup will change the skin color and may cover the blemishes totally. We find plausible makeup removal results can be achieved by decoding the base representation of with-makeup images only while setting the makeup representation to zero. That is, the makeup removal result can be generated by $G(0, R_b^y)$ for makeup image y. Two examples are shown in Figure 10, from which we can see the makeup of eyeshadow, eyelash and eyebrow have been removed successfully while the eye structure is preserved.

4.6. Many-to-Many Makeup Transfer

Our method can handle the makeup transfer between different eyes shapes, rotations, styles and etc. In Figure 11, we select three different source images (left column) and three different reference images (top row). Each row presents the results of the same source image wearing different makeups. Each column presents the results of different source images wearing the same makeup. The outputs (3×3 lower right) consistently preserve the identity of the source image and the style of the reference image. This capability is quite useful in real application. The user can virtually try different makeups on his own face conveniently as long as providing a reference image (*e.g.*, a picture of some famous star), and choose the favorite style.

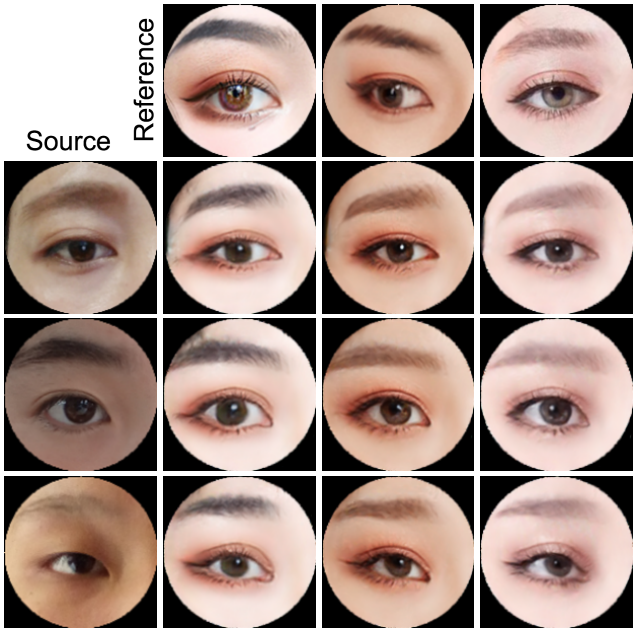

Figure 11: Various makeups of reference images are transferred to various source images.

5. Conclusion

In this paper, we incorporate novel semi-supervised learning, disentangling representation and autoencoder into an image domain adaption network. The eye makeup can be transferred from a reference face to a source face realistically and faithfully. Learned disentangled representations allow users to control the makeup degree by tuning makeup representation weight easily. Extensive experiments show that our method achieves state-of-the-art performance in transferring makeup styles. We have collected a dataset of various eye makeup styles, which will be made public. We believe this novel semi-supervised framework can be applied to other applications beyond makeup transfer and removal, which can be a future research direction.

Acknowledgments

This work was partly funded by MOE Tier-1 (RG132/18) project and MOE Tier-2 (2015-T2-2-114) Project.

References

[1] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018. 2

[2] E. Banijamali, A.-H. Karimi, A. Wong, and A. Ghodsi. Jade: Joint autoencoders for dis-entanglement. *arXiv preprint arXiv:1711.09163*, 2017. 2

[3] D. Bouchacourt, R. Tomioka, and S. Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[4] C. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in beta-vae. In *NIPS 2017 Disentanglement Workshop*, 2017. 2

[5] H. Chang, J. Lu, F. Yu, and A. Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 6, 7, 8

[6] T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018. 2

[7] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016. 2

[8] C. Donahue, A. Balsubramani, J. Mcauley, and Z. C. Lipton. Semantically decomposing the latent spaces of generative adversarial networks. *international conference on learning representations*, 2018. 2

[9] E. Dupont. Joint-vae: Learning disentangled joint continuous and discrete representations. *arXiv preprint arXiv:1804.00104*, 2018. 2

[10] Z. Feng, X. Wang, C. Ke, A. Zeng, D. Tao, and M. Song. Dual swap disentangling. *neural information processing systems*, pages 5894–5904, 2018. 2

[11] S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan. Auto-encoding total correlation explanation. *arXiv preprint arXiv:1802.05822*, 2018. 2

[12] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 2

[13] A. Gonzalezgarcia, J. V. De Weijer, and Y. Bengio. Image-to-image translation for cross-domain disentanglement. *neural information processing systems*, pages 1287–1298, 2018. 2

[14] D. Guo and T. Sim. Digital face makeup by example. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 73–79. IEEE, 2009. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017. 2

[17] X. Huang and S. J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, 2017. 5, 6

[18] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. *arXiv preprint arXiv:1804.04732*, 2018. 2, 3, 6, 7

[19] H. Kim and A. Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018. 2

[20] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, 4:3581–3589, 2014. 2

[21] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 6

[22] D.-T. Lee and B. J. Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980. 2, 4

[23] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1808.00948*, 2018. 2

[24] C. Li, K. Zhou, and S. Lin. Simulating makeup through physics-based manipulation of intrinsic image layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4621–4629, 2015. 2

[25] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 6, 7

[26] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 2

[27] S. Liu, X. Ou, R. Qian, W. Wang, and X. Cao. Makeup like a superstar: Deep localized makeup transfer network. *arXiv preprint arXiv:1604.07102*, 2016. 2

[28] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Scholkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *international conference on machine learning*, 2019. 2

[29] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 5

[30] S. Narayanaswamy, T. B. Paige, J.-W. van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr. Learning disentangled representations with semi-supervised deep generative models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5925–5935. Curran Associates, Inc., 2017. 2

[31] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 2

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[33] W.-S. Tong, C.-K. Tang, M. S. Brown, and Y.-Q. Xu. Example-based cosmetic transfer. In *Computer Graphics and Applications, 2007. PG'07. 15th Pacific Conference on*, pages 211–218. IEEE, 2007. 2

[34] C. Wang, C. Wang, C. Xu, and D. Tao. Tag disentangled generative adversarial network for object image re-rendering. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2901–2907, 2017. 2

[35] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 6

[36] Z. Yi, H. R. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2868–2876, 2017. 2

[37] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2, 3