Unsupervised Eyeglasses Removal in the Wild

Bingwen Hu, Zhedong Zheng, Ping Liu, *Member, IEEE*, Wankou Yang, and Mingwu Ren

***Abstract*—Eyeglasses removal is challenging in removing different kinds of eyeglasses, e.g., rimless glasses, full-rim glasses, and sunglasses, and recovering appropriate eyes. Due to the significant visual variants, the conventional methods lack scalability. Most existing works focus on the frontal face images in the controlled environment, such as the laboratory, and need to design specific systems for different eyeglass types. To address the limitation, we propose a unified eyeglass removal model called the eyeglasses removal generative adversarial network (ERGAN), which could handle different types of glasses in the wild. The proposed method does not depend on the dense annotation of eyeglasses location but benefits from the large-scale face images with weak annotations. Specifically, we study the two relevant tasks simultaneously, that is, removing eyeglasses and wearing eyeglasses. Given two face images with and without eyeglasses, the proposed model learns to swap the eye area in two faces. The generation mechanism focuses on the eye area and invades the difficulty of generating a new face. In the experiment, we show the proposed method achieves a competitive removal quality in terms of realism and diversity. Furthermore, we evaluate ERGAN on several subsequent tasks, such as face verification and facial expression recognition. The experiment shows that our method could serve as a preprocessing method for these tasks.**

***Index Terms*—Eyeglasses removal, generative adversarial network (GAN), image manipulation.**

I. Introduction

THE EYE is viewed as "a window to the soul" [1], containing rich biometric information, e.g., identity, gender, and age. In recent years, there are increasing interests in face-related applications. Among these applications, eyeglasses are usually considered as one kind of occlusion in the face images. As a result, the occlusion compromises downstream tasks, such as face verification [2]–[8] and expression recognition [9]–[11]. One way to address occlusion is by ignoring the occluded area, which is successfully applied in the field of person reidentification [12]–[16]. Different from the human body, the face is rich in identity information, and the eye area is the most discriminative facial field. The retained information is insufficient to support us making an accurate decision while the occluded area is ignored. Therefore, we propose an eyeglasses removal method to transform the occlusion area into a nonocclusion area. Despite significant advances in image manipulation, eyeglasses removal in the unconstrained environment, as known as in the wild, has not been well studied. In this article, we intend to fill this gap.

The previous eyeglasses removal works mainly focus on the cases in a controlled environment. Most works [17]–[21] require pairwise training samples. Every pair of images contains two frontal faces of the same identity with and without glasses. When testing, given one frontal testing image, the framework first detects the eye area and then replaces the original eye area with the reconstructed eye. The crafted eye area fuses the no-glasses training samples by principal component analysis (PCA) [22]. These lines of works adopt a strong assumption that faces are in a frontal pose, and the types of eyeglasses are limited. In the realistic scenario, however, there are three main drawbacks: 1) it is hard to collect a large number of pairwise training data of the same people with and without eyeglasses; 2) eyeglasses are usually made of different materials with significant visual variants, such as color, style, and transparency [see Fig. 1(a)]. It is infeasible to train dedicated removal systems for different eyeglasses types; and 3) existing works could not be applied to the face images in different poses. The model trained in the laboratory environment usually lacks scalability to significant visual variants.

This article addresses these three challenges. First, the previous methods [17]–[21] focused on the data collected in a controlled environment. More than that, they need the collected training images to be pairwised, that is, two frontal faces of the same identity with and without glasses. Other than that, some works [17]–[19] require an extra detector to locate the eyeglasses. Benefiting from the advancement of unsupervised learning [23], [24], we propose an unsupervised eyeglasses removal method. The only information that we need is whether the training image contains eyeglasses on the face. For training the proposed model in this article, we have collected 202 599 training images from the public dataset CelebA [25]. Those collected images are divided into two sets, that is, face images with glasses and face images without glasses, respectively.

Manuscript received February 11, 2020; accepted May 15, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61773117 and Grant 61703209, and in part by the National Major Scientific Instruments and Equipments Development Project of China under Grant 61727802. This article was recommended by Associate Editor X. Li. *(Corresponding author: Mingwu Ren.)*

Bingwen Hu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210092, China, and also with the Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: hubw.sky@gmail.com).

Zhedong Zheng is with the Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: zhedong.zheng@student.uts.edu.au).

Ping Liu was with the Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia. He is now with the Institute of High Performance Computing, A*STAR, Singapore (e-mail: pino.pingliu@gmail.com).

Wankou Yang is with the School of Automation, Southeast University, Nanjing 210092, China (e-mail: wkyang@seu.edu.cn).

Mingwu Ren is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210092, China (e-mail: renmingwu@njust.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2020.2995496

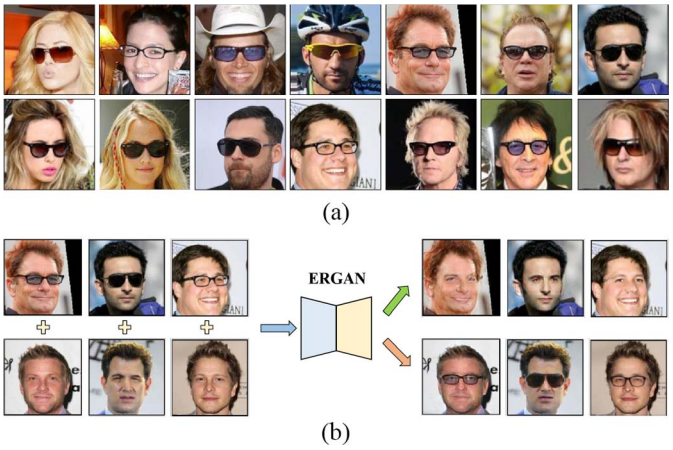


Fig. 1. (a) Examples of different types of eyeglasses in the wild. Different types of eyeglasses have significant visual variants, such as color, style, and transparency. Besides, the faces in the wild are usually in arbitrary poses with various lighting and backgrounds. (b) Brief pipeline of the proposed method. The proposed ERGAN simultaneously takes the two relevant tasks, i.e., wearing and removing glasses, into consideration.

This weak labeling demand for training sets largely saves on annotation cost.

Second, eyeglasses usually have significant intraclass variants in terms of geometry shape and appearance. It is infeasible to build dedicated models for every kind of eyeglasses. We, therefore, exploit an alternative method that lets the model "see" various eyeglasses and learn the scalable weights from a large number of face images. In particular, we propose the eyeglass removal GAN (ERGAN) to learn the general structure of glasses and encode different types of eyeglasses. The proposed method not only removes the eyeglasses but also has the capability to generate the eyeglasses, which further enforces the model to learn the local patterns of different eyeglasses [see Fig. 1(b)].

Third, the problem of eyeglasses removal from arbitrary face poses is common and challenging in a realistic scenario. The previous works mostly assume that we could obtain frontal face images and accurately align the eye area. In real scenarios, the user, however, may upload nonfrontal face images, and the eye area is not well aligned. One solution is to resort to alignment calibration to generate frontal faces, which might be complicated and time consuming. In contrast, the proposed method does not need alignment densely. The only preprocessing required is to rotate the face images according to the center of the face. Compared with previous works, the proposed method does not need complicated alignment and is robust to the various pose variants.

To address the above-mentioned challenges, we propose a unified ERGAN, which only needs weak annotations for training. We learn two representations of the input face image, that is, the face appearance code and the eye-area attribute code. The face appearance code mainly contains the low-level geometric structure of the face, while the eye-area attribute code captures the semantic pattern in the eye area. In more detail, we first utilize two different encoders to decompose the face image into a face appearance code and an eye-area attribute code, respectively. Then, one decoder is learned to combine the face appearance code with the eye-area attribute code to reconstruct the face image. The self-reconstruction loss is applied to ensure that the two latent codes complement each other and preserve the information of the original image. To enforce the model focuses on the eye area of the input image, we introduce the eye-area reconstruction loss. Furthermore, the face appearance reconstruction loss and the cycle consistency loss are adopted to encourage that the mapping function is invertible between the reconstructed face image and the two latent codes. When testing, we could combine the face appearance code of the target face with the eye-area attribute code to remove or wear any specific eyeglasses.

To evaluate the generation quality, we adopt the Fréchet inception distance (FID) [26] and learned perceptual image patch similarity (LPIPS) [27] as an indicator to test the realism and diversity of the generated face images, respectively. Extensive qualitative and quantitative experiments show that our method achieves superior performance to several existing approaches [28]–[30] and could serve as a preprocessing tool for subsequent tasks, such as face verification and facial expression recognition.

In summary, the main contributions of this article are as follows.

1) We propose a unified framework called ERGAN that enables different types of eyeglasses removal from faces in the wild. Compared with previous works, the proposed method does not need densely aligned faces nor pairwise training samples. The only weak annotation that we need is whether the training data contain eyeglasses or not.
2) Due to the large visual variants of the eyeglasses, including shape and color, eyeglasses removal demands to learn a robust model. In this article, we propose a dual learning scheme to simultaneously learn two inverse manipulations, that is, removing eyeglasses and wearing eyeglasses. Specifically, we utilize the eye-area reconstruction loss to explicitly make the model pay more attention to the eye area. The ablation study verifies the effectiveness of both the dual learning scheme and the eye-area reconstruction loss.
3) The qualitative and quantitative experiments show that our method outperforms other competitive approaches in terms of realism and diversity. Furthermore, we evaluate the proposed method on other face-related tasks. Attribute to the high-fidelity eyeglass removal, the generated results benefit the subsequent tasks, such as face verification and facial expression recognition.

II. Related Work

A. Statistical Learning

Most pioneering works on eyeglasses removal are based on statistical learning [17], [18], [20], [21], [31]. One line of works adopts the assumption that the target faces could be reconstructed from other faces without eyeglasses. Based on this assumption, previous methods widely adopt PCA [22] to learn the shared components among the face images. In one of the early works, Wu *et al.* [17] proposed a find-and-replace

approach to remove eyeglasses from frontal face images. This method first finds the location of eyeglasses by an eye-region detector and then replaces it with a synthesized glasses-free image. The synthesized glass-free image is inferred by combining the original eye area and different weighted glasses-free faces in the training data. Furthermore, Park *et al.* [18] applied the recursive process of PCA reconstruction and error compensation to generate facial images without glasses. Due to the different temperatures between glasses and human faces, another line of works takes advantage of thermal images to remove the eyeglasses. Wong and Zhao [21] proposed a nonlinear eyeglasses removing algorithm for thermal images based on kernel PCA (KPCA) [32]. This method performs KPCA to transfer the visible reconstruction information from the visible feature space to the thermal feature space, and then apply the image reconstruction to remove eyeglasses from the thermal face image. Different from the method based on PCA mentioned above, some researchers resort to sparse coding and expectation–maximization to reconstruct faces. Yi and Li [20] deployed the sparse representation technique in local feature space to deal with the issue of eyeglasses occlusion. De Smet *et al.* [31] proposed a generalized expectation–maximization approach, which applies the visibility map to inpaint the occluded areas of the face image.

However, these methods [17], [18], [20], [21], [31] are usually designed for frontal faces and specific eyeglasses in a controlled environment. Different from the existing work, our method focuses on face images collected from the realistic scenario, and is scalable to visual variants, such as pose, illumination, and different types of glasses.

B. Generative Adversarial Network

Recent advance in facial manipulation is due to two factors: 1) large-scale public face datasets with attribute annotations, for example, CelebA [25] and 2) the high-fidelity images generated by the generative adversarial network (GAN). GAN is one kind of the generative model benefiting from the competition between the generator and the discriminator [33]–[38]. The facial image manipulation algorithms based on GANs have taken significant steps [28]–[30], [39]–[41]. One line of previous work directly learns the image-to-image translation between different facial attributes. Shen and Liu [39] presented a GAN-based method using residual image learning and dual learning to manipulate the attribute-specific face area. Liu *et al.* [41] presented a unified selective transfer network for arbitrary image attribute editing (STGAN), by combining difference attribute vector and selective transfer unit (STU) in the autoencoder network. Another line of works first learn the embedding of the face attributes and then decode the learned feature to generate images. Liu *et al.* [29] proposed an unsupervised image-to-image translation (UNIT) framework, which combines the spirit of both GANs and variational autoencoders (VAEs) [42]. UNIT adopts the assumption that a pair of images in different domains can be mapped to a shared latent feature space, and thus, could conduct the face images translation by manipulating the latent code. Furthermore, Huang *et al.* [30] proposed a method for the multimodal UNIT (MUNIT). MUNIT assumes that a pair of images in different domains share the same content space but the style space. Sampling different style codes could produce diverse and multimodal outputs while preserving the principle content.

Compared with previous GAN-based approaches [28]–[30], the proposed ERGAN has significant differences listed as follows.

1) To the introduced eye-area loss and invertible eye generation, the proposed ERGAN explicitly focuses on the manipulation of the eye region. The conventional generation mechanism inevitably introduces the noise to other areas of the original face when removing glasses.
2) We adopt the instance generation mechanism, which could swap the eye area of any two facial images. Compared with the CycleGAN-based methods, the proposed method could generate more diverse images.
3) Different from the previous works treating the face with different attributes as two or multiple domains, we view the face images as one domain with two codes, that is, the face appearance code and the eye-area attribute code. By utilizing this fine-grained latent representation decomposition, we could manipulate the code to generate conditional outputs to meet user demands.

III. Methodology

We first formulate the problem of eyeglasses removal and provide an overview of the proposed ERGAN in Section III-A. In Section III-B, we describe the details of each component in ERGAN, followed by the full objective function in Section III-C.

A. Formulation

In this article, we assume each face image could be decomposed into a face appearance code and an eye-area attribute code. Based on the assumption, we combine a face appearance code with an eye-area attribute code from an example face image to remove or wear eyeglasses. In this case, we could handle two inverse manipulations at the same time, that is, eyeglasses removal and eyeglasses wearing. The architecture of the proposed unified ERGAN is presented in Fig. 2(a). We denote face images with glasses and without glasses as $\mathcal{X}$ and $\mathcal{Y}$ ($\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^{H\times W\times 3}$), respectively. The goal of the proposed method is to learn two mappings: 1) $\mathcal{X} \rightarrow \mathcal{Y}$ and 2) $\mathcal{Y} \rightarrow \mathcal{X}$ that could transfer an image $x \in \mathcal{X}$ to another image $y \in \mathcal{Y}$, and vice versa.

Generator: As illustrated in Fig. 2(a), the generator of the proposed method adopts a similar framework of autoencoder [43], [44], which consists of face appearance encoders ($E^f_{\mathcal{X}}, E^f_{\mathcal{Y}}$), eye-area attribute encoders ($E^e_{\mathcal{X}}, E^e_{\mathcal{Y}}$), and decoders ($G_{\mathcal{X}}, G_{\mathcal{Y}}$). Specifically, the role of encoders ($E^f_{\cdot}, E^e_{\cdot}$) is to encode a given face image into the face appearance code $f_{\cdot}$ and the eye-area attribute code $e_{\cdot}$, respectively. Moreover, $f_{\cdot}$ and $e_{\cdot}$ are combined to generate a new image with the decoder $G_{\cdot}$. The decoder $G_{\cdot}$ is a deterministic function and has inverse encoders $(E^f_{\cdot}, E^e_{\cdot}) = (G_{\cdot})^{-1}$. In particular, $G_{\mathcal{X}}$ generates face images without glasses and $G_{\mathcal{Y}}$ generates face images with glasses, respectively.

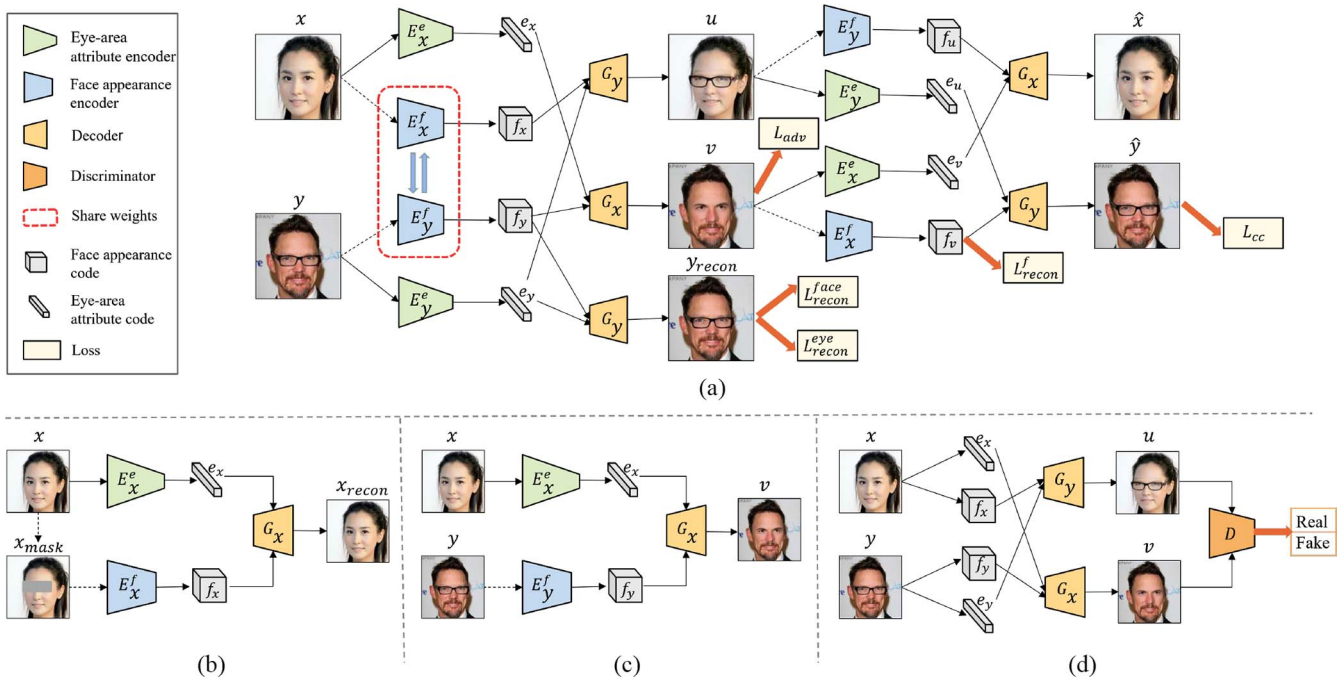


Fig. 2. Overview of ERGAN. (a) Proposed method implements two mappings: $\mathcal{X} \rightarrow \mathcal{Y}$ and $\mathcal{Y} \rightarrow \mathcal{X}$. The input face image is decomposed to a face appearance code f and an eye-area attribute code e by encoders $E^f_{\cdot}$ and $E^e_{\cdot}$, respectively. Encoders $E^f_{\mathcal{X}}$ and $E^f_{\mathcal{Y}}$ share weights. The black dashed line denotes that the eye region of the input image to face appearance encoder $E^f_{\cdot}$ is masked. (b) Self-reconstruction process of input image x to generate x_{recon}. x_{mask} denotes x after the eye area is masked. (c) Illustration of example-guided eyeglasses removal. e_x and f_y are combined to generate v by $G_{\mathcal{X}}$. (d) We propose a dual learning scheme to simultaneously learn two inverse manipulations (removing eyeglasses and wearing eyeglasses) by swapping eye-area attribute codes.

Discriminator: $(D_{\mathcal{X}}, D_{\mathcal{Y}})$ are two discriminators for $\mathcal{X}$ and $\mathcal{Y}$, respectively. The discriminator $D_{\cdot}$ aims to distinguish between generated images and real images. For instance, given a pair of images $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the discriminator $D_{\mathcal{X}}$ aims to distinguish images generated by decoder $G_{\mathcal{X}}(E^f_{\mathcal{X}}(x), E^e_{\mathcal{Y}}(y))$ from real images in $\mathcal{X}$. In the same way, the discriminator $D_{\mathcal{Y}}$ aims to distinguish images generated by decoder $G_{\mathcal{Y}}(E^f_{\mathcal{Y}}(x), E^e_{\mathcal{X}}(x))$ from real images in $\mathcal{Y}$. Specifically, the generated image by decoder $G_{\mathcal{X}}(E^f_{\mathcal{X}}(x), E^e_{\mathcal{Y}}(y))$ has the same face appearance code of x and the same eye-area attribute code of y. In contrast, the generated image by decoder $G_{\mathcal{Y}}(E^f_{\mathcal{Y}}(x), E^e_{\mathcal{X}}(x))$ has the same face appearance code of y and the same eye-area attribute code of x.

B. Eyeglasses Removal Generative Adversarial Network

Face Image Self-Reconstruction: As shown in Fig. 2(b), to enforce the generator focuses on the eye area of the input face image x, we first mask out the eye region to generate x_{mask} and then encode x and x_{mask} by encoders $(E^f_{\mathcal{X}}, E^e_{\mathcal{X}})$ to obtain the eye-area attribute code e_x and the face appearance code f_x. Finally, e_x and f_x are combined to generate the self-reconstructed image x_{recon} with the decoder $G_{\mathcal{X}}$, where $x_{\text{recon}} = G_{\mathcal{X}}(E^f_{\mathcal{X}}(x), E^e_{\mathcal{X}}(x))$. It is straightforward that the generated result of self-reconstruction approximates the source image. We introduce the face self-reconstruction loss, which is defined as

$$L^{\text{face}}_{\text{recon}} = \mathbb{E}\Big[\Big\| G_{\mathcal{X}}\Big(E^f_{\mathcal{X}}(x), E^e_{\mathcal{X}}(x)\Big) - x\Big\|_1\Big] + \mathbb{E}\Big[\Big\| G_{\mathcal{Y}}\Big(E^f_{\mathcal{Y}}(y), E^e_{\mathcal{Y}}(y)\Big) - y\Big\|_1\Big] \quad (1)$$

where $\mathcal{X}$ represents the set of face images without glasses and $\mathcal{Y}$ represents the set of face images with glasses, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The pixelwise ℓ_1-norm $\|\cdot\|_1$ is employed for preserving the sharpness of self-reconstruction images. The face self-reconstruction loss enforces the encoders $(E^f_{\mathcal{X}}, E^e_{\mathcal{X}})$ to learn two representations of the input face image, that is, the face appearance code and the eye-area attribute code, respectively.

Eye-Area Reconstruction: The face image self-reconstruction loss encourages the model to focus on the global features of the input image. For the glasses removal task, employing only the face image self-reconstruction loss misses the specific information of the eye area of the generated image. Therefore, enforcing the model to pay more attention to the eye area, we introduce the eye-area reconstruction loss

$$L^{\text{eye}}_{\text{recon}} = \mathbb{E}\big[\big\| x^{\text{eye}}_{\text{recon}} - x^{\text{eye}}\big\|_1\big] + \mathbb{E}\big[\big\| y^{\text{eye}}_{\text{recon}} - y^{\text{eye}}\big\|_1\big] \quad (2)$$

where x^{eye} and $x^{\text{eye}}_{\text{recon}}$ are denoted as the eye area of x and x_{recon}, respectively. Similarly, y^{eye} and $y^{\text{eye}}_{\text{recon}}$ are denoted as the eye area of y and y_{recon}. In practice, since the face images are all center-aligned, the eye area is defined as $(x_1 = 0.4h, x_2 = 0.2w, y_1 = 0.65h, y_2 = 0.75w)$ area of the input image, where (h, w) is the size of the input face image.

Dual Learning Scheme: The proposed method is based on the assumption that the face image could be decomposed into a face appearance code and an eye-area attribute code. In detail, given a face appearance code, the decoder $G_{\cdot}$ could combine the face appearance code with the eye-area attribute code from the target face image to generate an image with or without glasses. Fig. 2(c) shows an example of eyeglasses removal.

The two related tasks of removing glasses and wearing glasses could be regarded as a dual process. Therefore, we apply a dual learning scheme [45] to learn two inverse manipulations simultaneously. The ablation study (in Section IV-F) confirms the effectiveness of the dual learning scheme.

We show the process of dual learning in Fig. 2(d). For given images x and y, we encode them into (f_x, e_x) and (f_y, e_y), where $(f_x, e_x) = (E^f_{\mathcal{X}}, E^e_{\mathcal{X}})$ and $(f_y, e_y) = (E^f_{\mathcal{Y}}, E^e_{\mathcal{Y}})$. The dual tasks (i.e., removing eyeglasses and wearing eyeglasses) are performed by swapping the eye-area attribute codes. We adopt decoders $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$ to generate the final output images $u = G_{\mathcal{Y}}(f_x, e_y)$ and $v = G_{\mathcal{X}}(f_y, e_x)$, respectively. In particular, the decoder $G_{\mathcal{Y}}$ combines the face appearance code of x and the eye-area attribute code of y to generate u. Similarly, the decoder $G_{\mathcal{X}}$ combines the face appearance code of y and the eye-area attribute code of x to generate v.

Based on the assumption that each face image could be decomposed into a face appearance code and an eye-area attribute code, in this article, we learn two representations of the input face image, that is, the face appearance code and the eye-area attribute code. The face appearance code mainly contains the low-level geometric structure of the face. Since the encoder maps the region outside the eye area to the face appearance space is irrelevant to the face image with or without glasses. During the training, therefore, we employ the weights sharing between $E^f_{\mathcal{X}}$ and $E^f_{\mathcal{Y}}$ to update the model synchronously. In contrast, the eye-area attribute code captures the semantic pattern in the eye area, resulting in the eye-area attributes of face images with glasses or without glasses are significantly different. In this case, we do not share weights between $E^e_{\mathcal{X}}$ and $E^e_{\mathcal{Y}}$.

Furthermore, the proposed method should be able to reconstruct (f_x, e_x) and (f_x, e_y) after decoding u and v. As illustrated in Fig. 2(d), we apply encoders $E^f_{\mathcal{X}}$ and $E^e_{\mathcal{Y}}$ obtain the face appearance code f_x and the eye-area attribute code e_y, then f_x and e_y are concatenated together to generate u. A similar process is utilized to generate v. Then, we encode u and v into (f_u, e_u) and (f_v, e_v). To ensure that generated images u and v retain the original information, we define the face appearance reconstruction loss L^f_{recon} and the eye-area attribute reconstruction loss L^e_{recon}. We formulate L^f_{recon} and L^e_{recon} as follows:

$$L^f_{\text{recon}} = \mathbb{E}\big[\|f_u - f_x\|_1\big] + \mathbb{E}\big[\|f_v - f_y\|_1\big] \tag{3}$$

$$L^e_{\text{recon}} = \mathbb{E}[\|e_u - e_x\|_1] + \mathbb{E}\big[\|e_v - e_y\|_1\big] \tag{4}$$

where $(f_u, f_v) = (E^f_{\mathcal{X}}(u), E^f_{\mathcal{Y}}(v))$ and $(e_u, e_v) = (E^e_{\mathcal{X}}(u), E^e_{\mathcal{Y}}(v))$. Notably, we find that the performance of the model is declined by introducing the eye-area attribute reconstruction loss L^e_{recon}. To achieve the highest performance, we ignore L^e_{recon}. The detailed discussion in our ablation study (in Section IV-F).

Then, we recombine face appearance codes (f_u, f_v) and eye-area attribute codes (e_u, e_v) to generate $\hat{x} = G_{\mathcal{X}}(f_u, e_v) = G_{\mathcal{X}}(E^f_{\mathcal{X}}(u), E^e_{\mathcal{Y}}(v))$ and $\hat{y} = G_{\mathcal{Y}}(f_v, e_u) = G_{\mathcal{Y}}(E^f_{\mathcal{Y}}(v), E^e_{\mathcal{X}}(u))$, respectively. To ensure the generated images u and v reconstruct the original images x and y, we introduce the cycle consistency loss [28]. The cycle consistency loss L_{cc} is defined as

$$\begin{aligned} L_{cc} = {} & \mathbb{E}\Big[\Big\|G_{\mathcal{X}}\Big(E^f_{\mathcal{X}}(u), E^e_{\mathcal{Y}}(v)\Big) - x\Big\|_1\Big] \\ & + \mathbb{E}\Big[\Big\|G_{\mathcal{Y}}\Big(E^f_{\mathcal{Y}}(v), E^e_{\mathcal{X}}(u)\Big) - y\Big\|_1\Big] \end{aligned} \tag{5}$$

where $u = G_{\mathcal{Y}}(f_x, e_y)$ and $v = G_{\mathcal{X}}(f_y, e_x)$.

Adversarial Loss: To encourage the generated face image indistinguishable from the real face image, we adopt the adversarial loss [33]. In this case, $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$ attempt to generate high-fidelity face images (e.g., with glasses and without glasses), while $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ attempt to distinguish real face images from generated face images. The adversarial loss is defined as

$$\begin{aligned} L_{\text{adv}} = {} & \mathbb{E}\big[\log D_{\mathcal{X}}(x)\big] + \mathbb{E}\big[\log\big(1 - D_{\mathcal{X}}\big(G_{\mathcal{X}}\big(f_y, e_x\big)\big)\big)\big] \\ & + \mathbb{E}\big[\log D_{\mathcal{Y}}(y)\big] + \mathbb{E}\big[\log\big(1 - D_{\mathcal{Y}}\big(G_{\mathcal{Y}}\big(f_x, e_y\big)\big)\big)\big]. \end{aligned} \tag{6}$$

Discussion: Different from several image-to-image translation frameworks [28]–[30] to manipulate the whole image, the proposed method focuses on the manipulation of the eye area, which does not change the regions outside the eye area and preserves the information of the original image effectively. Specifically, inspired by image inpainting works [46]–[48], we cover the eye area of the input face image and then apply the encoder $E^f_{\cdot}$ to obtain the face appearance code, which avoids the interference of original eye area information. Moreover, we use the encoder $E^e_{\cdot}$ to obtain the eye-area attribute code of the input face image. In particular, introducing both the face image self-reconstruction loss and the eye-area reconstruction loss enforces the model to learn two representations of the input image, that is, the face appearance code and the eye-area attribute code. Besides, both the face image self-reconstruction loss and the eye-area reconstruction loss enforce the proposed model only to manipulate the eye region while maintaining the rest regions unchanged. Furthermore, we argue that the information feedback mechanism by dual learning could effectively improve the performance of the proposed method. Accordingly, we apply the dual learning scheme to realize the two inverse tasks of wearing glasses and removing glasses. The ablation study (in Section IV-F) demonstrates the effectiveness of the dual learning scheme.

C. Objective Function

Taking all above loss functions into account, we jointly train the encoders, decoders, and discriminators. We formulate the full objective function as

$$\begin{aligned} & \min_{E^e_{\cdot}, E^f_{\cdot}, G_{\mathcal{X}}, G_{\mathcal{Y}}} \max_{D_{\mathcal{X}}, D_{\mathcal{Y}}} L_{\text{total}}\Big(E^e_{\cdot}, E^f_{\cdot}, G_{\mathcal{X}}, G_{\mathcal{Y}}, D_{\mathcal{X}}, D_{\mathcal{Y}}\Big) \\ & = \lambda_{\text{face}} L^{\text{face}}_{\text{recon}} + \lambda_{\text{eye}} L^{\text{eye}}_{\text{recon}} + L^f_{\text{recon}} + L_{cc} + L_{\text{adv}} \end{aligned} \tag{7}$$

where the hyperparameters λ_{face} and λ_{eye} control the weights of the face self-reconstruction loss and the eye-area reconstruction loss.

IV. Experiment

A. Datasets

CelebA: The CelebA dataset [25] consists of 202 599 aligned face images collected from 10 177 celebrities in

the wild. Each image in CelebA is annotated with 40 face attributes (e.g., with/without eyeglasses and smiling/no-smiling). We split the CelebA dataset into one subset with glasses and another without glasses, based on the annotated attributes. Accordingly, we obtain 13 193 images with glasses and 189 406 images without glasses. All face images are center cropped to 160×160 and resized to 224×224 with a probability of 0.5 horizontal flipping.

LFW: The LFW dataset [49] contains 13 233 face images of 5749 identities collected from the uncontrolled surroundings. All images are aligned by deep funneling [50]. We choose 1600 face images with glasses and 8000 face images without glasses from the LFW dataset to verify the effectiveness of the proposed method. The preprocessing of the LFW dataset is the same as the CelebA dataset.

MeGlass: The MeGlass dataset [51] includes 47 917 face images of 1710 identities selected and cleaned from the MegaFace dataset [52]. The MeGlass dataset contains 14 832 face images with glasses and 33 087 face images without glasses. Besides, each identity contains at least two face images with glasses and two face images without glasses. We also perform experiments on MeGlass to verify the effectiveness of our method.

B. Evaluation Metrics

Fréchet Inception Distance: Most image generation tasks [53]–[56] adopt the FID metric [26], which is utilized to measure the distance between generated images and real images through feature extracted by the inception network [57]. In this article, we apply the FID metric to measure the realism of the generated face images.

Learned Perceptual Image Patch Similarity: The LPIPS [27] distance is used to evaluate the diversity of generated images. Following several state-of-the-art image-to-image translation approaches [30], [55], [58], [59], we employ the LPIPS distance to measure the diversity of manipulated face images. Because the proposed method focuses on the task of eyeglasses removal, we only capture the eye area to measure the LPIPS distance to make a fair comparison. We denote the modified LPIPS metric as eLPIPS.

C. Implementation Details

Here, we provide details about the network architecture of ERGAN.

1) The eye-area attribute encoder $E_\cdot^e$ consists of several convolutional layers and residual blocks [60] as well as a global average pooling layer and a fully connected layer.
2) For the face appearance encoder $E_\cdot^f$, it only consists of several convolutional layers and residual blocks.
3) The decoder $G_\cdot$ combines $E_\cdot^e$ and $E_\cdot^f$ through four residual blocks followed by several convolutional layers and upsampling layers. In particular, we finally append a refine block which consists of a convolutional layer and a fully connected layer. The refine block is used to concatenate the reconstructed image with the input image to produce a higher quality generated image. In addition, all convolutional layers are followed by instance normalization [61]. Similar to MUNIT [30], each residual block contains two adaptive instance normalization layers [62].
4) For the discriminator D, we apply the multiscale discriminator (PatchGAN) proposes by Wang *et al.* [63]. Moreover, we adopt the Leaky ReLU with slope 0.2 in both generator and discriminator. In particular, we show detailed network architectures of $E_\cdot^e$, $E_\cdot^f$, and $G_\cdot$ in Tables I–III.

TABLE I
NETWORK ARCHITECTURE OF THE EYE-AREA ATTRIBUTE ENCODER $E_\cdot^e$

Layer	Parameters	Output Size
Input	-	$3 \times 224 \times 224$
Conv1	$[\ 3\times3,\ 64\ ]$	$64 \times 224 \times 224$
Conv2	$[\ 3\times3,\ 64\ ]$	$64 \times 224 \times 224$
Conv3	$[\ 3\times3,\ 128\ ]$	$128 \times 112 \times 112$
ResBlocks	$\begin{bmatrix} 3\times3,\ 128 \\ 3\times3,\ 128 \end{bmatrix}\times2$	$128 \times 112 \times 112$
Conv4	$[\ 3\times3,\ 256\ ]$	$256 \times 56 \times 56$
ResBlocks	$\begin{bmatrix} 3\times3,\ 256 \\ 3\times3,\ 256 \end{bmatrix}\times2$	$256 \times 56 \times 56$
Conv5	$[\ 3\times3,\ 256\ ]$	$256 \times 28 \times 28$
Conv6	$[\ 3\times3,\ 256\ ]$	$256 \times 14 \times 14$
Conv7	$[\ 1\times1,\ 256\ ]$	$256 \times 1 \times 1$

TABLE II
NETWORK ARCHITECTURE OF THE FACE APPEARANCE ENCODER $E_\cdot^f$

Layer	Parameters	Output Size
Input	-	$3 \times 224 \times 224$
Conv1	$[\ 3\times3,\ 64\ ]$	$64 \times 224 \times 224$
Conv2	$[\ 3\times3,\ 64\ ]$	$64 \times 224 \times 224$
Conv3	$[\ 4\times4,\ 128\ ]$	$128 \times 112 \times 112$
ResBlocks	$\begin{bmatrix} 3\times3,\ 128 \\ 3\times3,\ 128 \end{bmatrix}\times2$	$128 \times 112 \times 112$
Conv4	$[\ 4\times4,\ 256\ ]$	$256 \times 56 \times 56$
ResBlocks	$\begin{bmatrix} 3\times3,\ 256 \\ 3\times3,\ 256 \end{bmatrix}\times2$	$256 \times 56 \times 56$

During the training, we adopt Adam optimizer [64] to optimize the generator and the discriminators. In addition, we set the initial learning rate to 0.0001, weight decay 0.0005, and exponential decay rates $(\beta_1, \beta_2) = (0, 0.999)$. Following several typical image-to-image translations [29], [30], [59], we set hyperparameters $\lambda_{\text{face}} = 10$ for the face self-reconstruction loss. To encourage the model to focus on the eye region, we set a large weight of $\lambda_{\text{eye}} = 10$ for the eye-area reconstruction loss. For the adversarial loss L_{adv}, we employ the LSGAN objective proposed by Mao *et al.* [65]. Moreover, the gradient punishment strategy [66] is also adopted to stabilize our model training procedure.

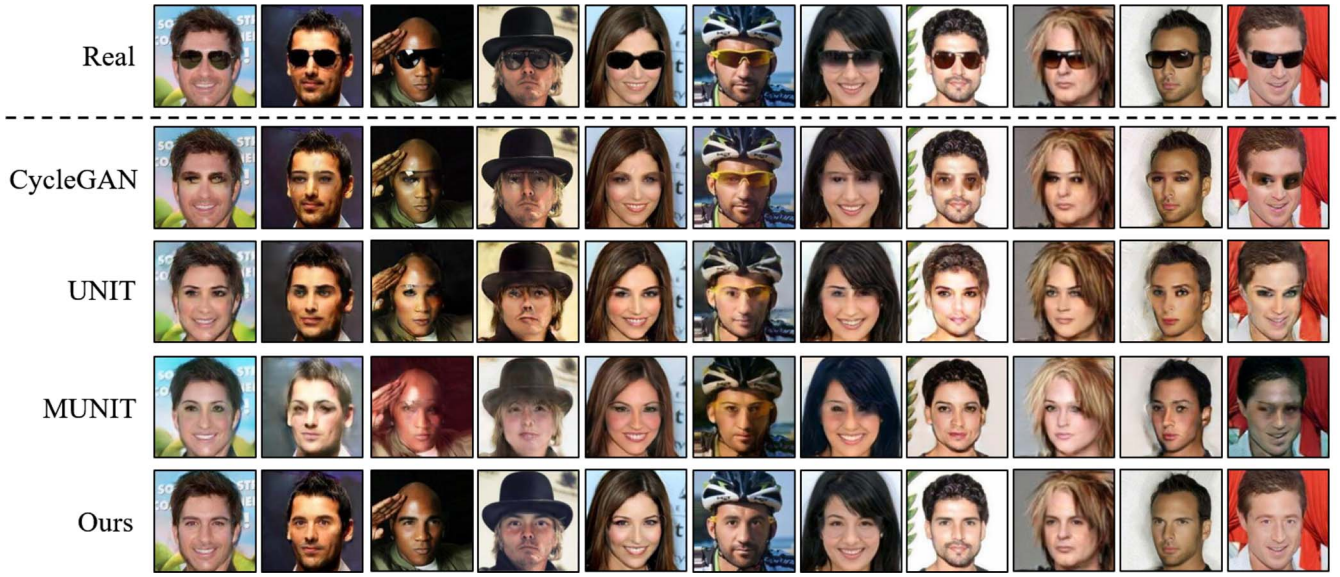


Fig. 3. Results of eyeglasses removal from face images on the CelebA dataset. From top to bottom: real images, CycleGAN [28], UNIT [29], MUNIT [30], and our method.

TABLE III
NETWORK ARCHITECTURE OF THE DECODER $G_{\cdot}$

Layer	Parameters	Output Size
Input	-	256 × 56 × 56
ResBlocks	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times4$	256 × 56 × 56
Unsample	-	256 × 112 × 112
Conv1	[3×3, 128]	128 × 112 × 112
Unsample	-	128 × 224 × 224
Conv2	[3×3, 64]	64 × 224 × 224
Conv3	[3×3, 64]	64 × 224 × 224
Conv4	[3×3, 3]	3 × 224 × 224
Refine Block	$\begin{bmatrix} 3\times3, 3 \\ 1\times1, 3 \end{bmatrix}\times1$	3 × 224 × 224

D. Competitive Methods

We compare the proposed method with several two-domain image-to-image translation frameworks, including CycleGAN [28], UNIT [29], and MUNIT [30]. To make a fair comparison, we train our method and [28]–[30] under the same setting.

CycleGAN [28]: CycleGAN combines adversarial loss and cycle consistency loss to train two GANs for image-to-image translation.

UNIT [29]: The UNIT model assumes that images of different domains could be mapped to the same latent representation, and the images generated by the model could be associated with input images of different domains by VAEs [42], respectively.

MUNIT [30]: The MUNIT model is a multimodal UNIT framework. It assumes that images from different domains could be decomposed into a shared content space and a style specific space. Therefore, an image could be translated to the target domain by recombining the content code of the image with a random style code in the target style space.

These competitive methods usually treat the face with different attributes as two or multiple domains. In contrast, we view the face images as one domain with two codes, that is, the face appearance code and the eye-area attribute code. Besides, these three competitive methods manipulate the whole image to remove glasses, while our method only operates the eye area and other regions remain unchanged.

E. Evaluations

Qualitative Evaluation: We first qualitatively compare our method with three generative approaches above mentioned. As shown in Figs. 3 and 4, we evaluate the generated results quality after eyeglasses removal from face images on the CelebA dataset and the LFW dataset, respectively. We reimplement the competitive methods, that is, CycleGAN [28], UNIT [29], and MUNIT [30] by the open-source code. As shown in Figs. 3 and 4, CycleGAN, UNIT, and MUNIT still have limitations in the manipulation of eyeglasses removal. These competitive methods are prone to generate blurry or oversmoothing results and remove glasses insufficiently. In comparison, our generated images are natural and realistic, suggesting that the proposed method is effective in removing eyeglasses from face images. In particular, as illustrated in Fig. 5, the manipulation of eyeglasses removal by three competitive methods also change rest regions outside the eye area, which dramatically reduces the quality of the generated image. In contrast, our method only manipulates the eye region while keeping the rest regions unchanged.

We show example-guided eye-area attribute manipulation results in Fig. 6. We observe that our method achieves high-quality results in two inverse operations, that is, removing glasses and wearing glasses. Moreover, the eye-area feature from the target image maintains faithfully. We further perform linear interpolation between two eye-area attribute codes

Fig. 4. Results of eyeglasses removal from face images on the LFW dataset. From top to bottom: real images, CycleGAN [28], UNIT [29], MUNIT [30], and our method.

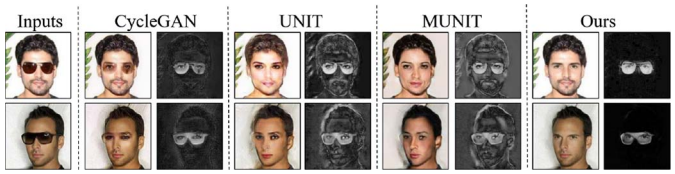


Fig. 5. Residual images obtained by our method and three competitive methods, e.g., CycleGAN [28], UNIT [29], and MUNIT [30].

to generate the corresponding face images, as shown in Fig. 7. The linear interpolation results demonstrate that the eye-area attribute of face images change smoothly with latent codes.

Quantitative Evaluation: Here, we report the results of quantitative evaluations based on FID and eLPIPS metrics aiming to measure the realism and the diversity of our generated face images. As shown in Table IV, our method obtains the minimum FID and the maximum eLPIPS on the CelebA dataset, suggesting that generated face images by our method have a close distribution of real face images. We proceed to perform our method on the LFW dataset. As shown in Table V, our method also achieves the minimum FID and the maximum eLPIPS. The result indicates that the proposed method has scalability compared to three competitive methods. Moreover, the quantitative evaluation verifies the authenticity of qualitative evaluation, and our method is significantly superior to three competitive methods on both realism and diversity.

F. Ablation Study

To study the contribution of each component in the proposed method, we perform several variants of ERGAN and evaluate them on the CelebA dataset. Specifically, we evaluate six variants, that is, 1) *Ours w/o* $L_{\text{recon}}^{\text{face}}$: our method without the face self-reconstruction loss term; 2) *Ours w/o* $L_{\text{recon}}^{\text{eye}}$: our method without the eye-area reconstruction loss term; 3) *Ours w/o* L_{recon}^{f}: our method without the face appearance reconstruction loss term; 4) *Ours w/o Lcc*: our method without the cycle consistency loss term; 5) *Ours* (*half*): our method without the dual learning scheme and only implementing the task of eyeglasses removal; and 6) *Ours w/* L_{recon}^{e}: our method with the eye-area attribute reconstruction loss term.

We show the qualitative results of six variants in Fig. 8. Without using the face self-reconstruction loss, our method can still generate plausible results. However, some noise is introduced into the generated results, suggesting that the face self-reconstruction loss is a critical factor in keeping the rest regions except the eye area remains unchanged. Without adopting the face appearance reconstruction loss and the cycle consistency loss, our method generates slightly distorted results. The results cannot preserve the consistency of the input image. Without employing the dual learning scheme and only implementing the task of eyeglasses removal, the model produces much lower quality results, demonstrating that the dual learning scheme is effective in learning features. We also evaluate the variant of our method with the eye-area attribute reconstruction loss. The results show that adopting the eye attribute reconstruction loss cannot improve the quality of generated images. We suspect that this constraint is too strong, increasing the interferences in such a small region of the eye area. To achieve the best performance of our method, we remove the eye attribute reconstruction loss.

We report the quantitative ablation study results of six variants in Table IV. It can be observed that the full method obtains a lower FID score than six variants, suggesting that the proposed method can generate more realistic images on two tasks. For diversity, the scores drop dramatically without applying the face self-reconstruction loss, which indicates that this constraint is the key component to generate diversity outputs. In comparison to the variant which only implements the single task of eyeglasses removal, our method achieves lower FID scores and higher eLPIPS scores. It demonstrates that adopting the dual learning scheme can generate higher quality

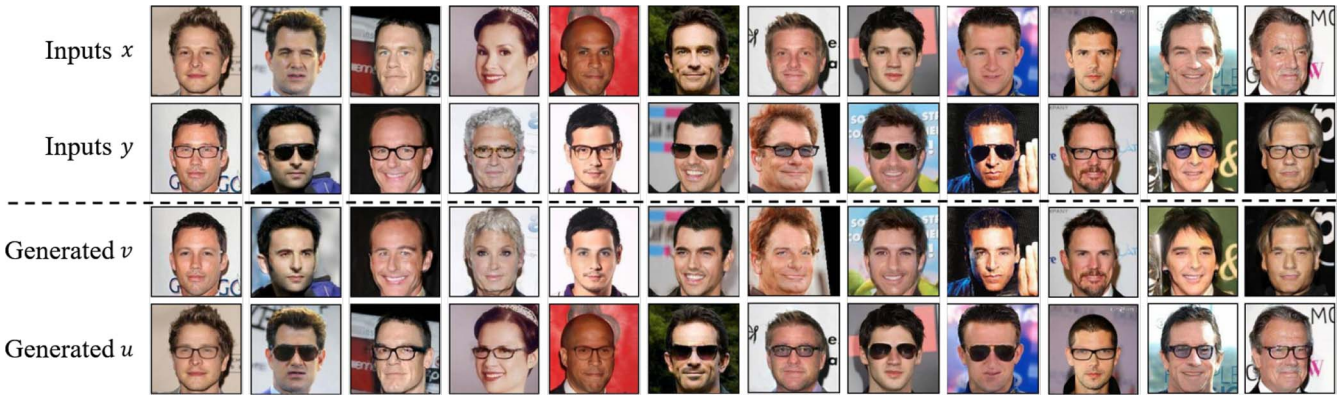


Fig. 6. Examples of our generated images by swapping eye-area attribute codes on the CelebA dataset.

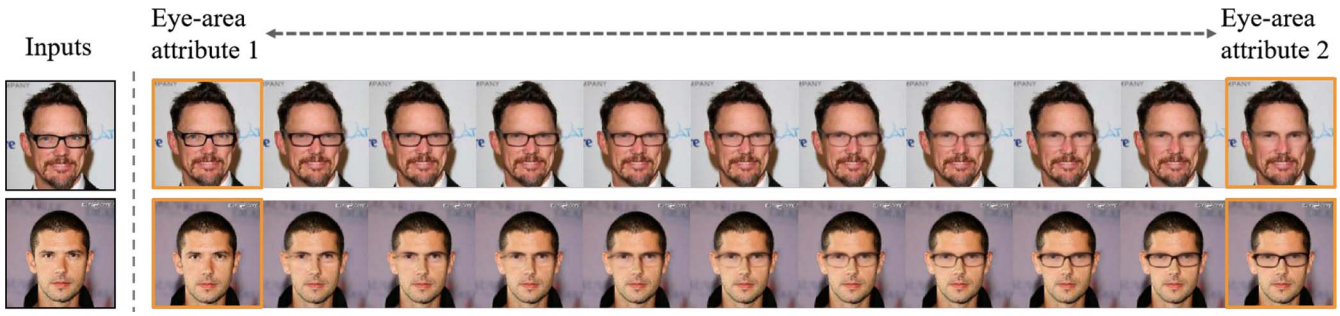


Fig. 7. Linear interpolation. Image generation results with linear interpolation between two eye-area attribute codes.

TABLE IV
QUANTITATIVE RESULTS. COMPARISON OF FID (LOWER IS BETTER) AND eLPIPS (HIGHER IS BETTER) TO EVALUATE REALISM AND DIVERSITY OF GENERATED FACE IMAGES AND THE REAL DATA ON THE CELEBA DATASET

Method	FID ↓		eLPIPS ↑	
	Wearing	Removal	Wearing	Removal
Real data	6.02	5.62	-	-
CycleGAN [28]	15.65	20.67	-	-
UNIT [29]	18.80	18.86	0.114	0.074
MUNIT [30]	29.42	18.85	0.283	0.144
Ours w/o L_{recon}^{face}	14.12	16.60	0.005	0.002
Ours w/o L_{recon}^{eye}	12.79	16.50	0.384	0.162
Ours w/o L_{recon}^{f}	14.42	15.68	0.432	0.208
Ours w/o L_{cc}	12.46	16.27	0.426	0.234
Ours (*half*)	-	15.87	-	0.010
Ours w/ L_{recon}^{e}	12.96	16.33	**0.435**	0.220
Ours (*full*)	**11.96**	**15.07**	0.428	**0.240**

TABLE V
QUANTITATIVE RESULTS. COMPARISON OF FID (LOWER IS BETTER) AND eLPIPS (HIGHER IS BETTER) TO EVALUATE REALISM AND DIVERSITY OF GENERATED FACE IMAGES AND THE REAL DATA ON THE LFW DATASET

Method	FID ↓		eLPIPS ↑	
	Wearing	Removal	Wearing	Removal
Real data	24.17	6.96	-	-
CycleGAN [28]	40.37	23.33	-	-
UNIT [29]	40.30	35.31	0.132	0.076
MUNIT [30]	51.74	42.83	0.201	0.141
Ours (*full*)	**26.58**	**19.87**	**0.367**	**0.260**

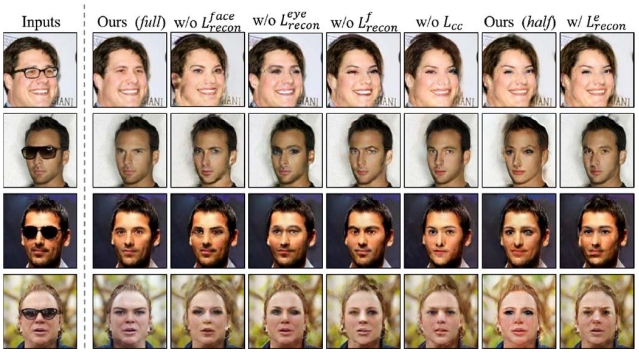


Fig. 8. Qualitative ablation study. Visualization results of six variants.

images in terms of realism and diversity. We observe that our method introduces the eye-area attribute reconstruction loss performs slightly better on glasses wearing in the diversity metric. The proposed method achieves the best performance on all the other indications, especially on eyeglasses removal in the realism metric.

G. Further Analysis and Discussion

Limitation: We notice that the proposed method tends to produce low-fidelity results when the input face image pairs have large pose changes. As shown in Fig. 9, given two face images with a similar pose, our method could generate high-quality results. In contrast, the quality of generated images is

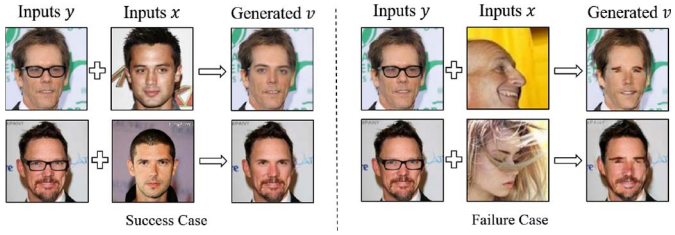


Fig. 9. Success and failure cases of eyeglasses removal applying our method.

degraded when the input images have significant pose changes (e.g., one front face and one profile face). There are two main reasons for the limitation. One is the limited amount of training data. Therefore, it is hard for our method to learn the two representations, that is, the face appearance code and the eye-area attribute code, to cover all the variations introduced by poses. Another reason is that the profile face usually contains noises, such as background and hairs, compromising the eye-area attribute code learning. As a result, the generated face is hard to simulate the target eye.

Face Verification: To further evaluate the performance of the proposed method, we first test whether it is beneficial to the face verification task. We apply the off-the-shelf face verification model FaceNet[1] [2]. Three test subsets of the LFW dataset are taken into consideration, that is, original test images $P1$, selected images with eyeglasses $P2$, and selected images removed glasses by the proposed method $P2_r$. $P1$ is the standard test set of the LFW dataset. Due to the limited number of face images with eyeglasses in LFW, we further sample a subset $P2$ from the original test set, which contains both matched pairs (with the same identity, one face image with glasses and another without glasses) and mismatched pairs (with different identities, both face images with glasses). The subset $P2_r$ contains the same images as $P2$, while those images are manipulated with glasses removal by our method.

We report quantitative face verification results on LFW in Table VI. Comparing the face verification results obtained on $P1$ and $P2$, we observe that the accuracy obtained on the test set $P2$ is lower than that obtained on $P1$. This result shows that the occlusion of glasses compromises the accuracy of the face verification model. However, this result does not indicate that the occlusion of glasses is the only factor that reduces accuracy. Other factors, including illumination, resolution, and viewpoint, still account for the reduced accuracy of face verification. Moreover, comparing the results between $P2$ and $P2_r$, we find that the performance of face verification improved after eyeglasses removal by the proposed method. Notably, it is reasonable to obtain a gain of 0.32% for face verification. Our method can alleviate the impact of glasses occlusion on face verification. Although other factors, such as illumination, resolution, and viewpoint exist and affect the accuracy of face verification, our method can improve the accuracy to a sensible extent.

For face verification, we also perform experiments on three subsets of the MeGlass dataset, that is, $M1$ that contains face images without glasses, $M2$ that contains face images with glasses, and $M2_r$ that includes the same samples as $M2$. $M2_r$ includes all images that have been manipulated with eyeglasses removal by our method. We report quantitative face verification results on MeGlass in Table VII. We find that the face verification accuracy obtained on the test set $M2_r$ is higher than the accuracy on $M2$. The result is consistent with our statement that eyeglass removal could help face verification. Therefore, the proposed method could improve the accuracy of face verification on the large-scale dataset.

TABLE VI
FQUANTITATIVE FACE VERIFICATION RESULTS ON LFW. COMPARISON OF THE FACE VERIFICATION ACCURACY ON $P1$, $P2$, AND $P2_r$ TO EVALUATE THE EFFECTIVENESS OF EYEGLASSES REMOVAL. $P2$ CONTAINS FACE IMAGES WITH GLASSES WHICH SELECT FROM ORIGINAL IMAGES. $P2_r$ CONTAINS THE SAME IMAGES WITH $P2$ WHILE THOSE IMAGES ARE MANIPULATED WITH GLASSES REMOVAL ADOPTING OUR METHOD

Face images	Accuracy
Original test images $P1$	97.80%
Selected images $P2$	96.35%
Selected images after eyeglasses removal $P2_r$	96.67%

TABLE VII
QUANTITATIVE FACE VERIFICATION RESULTS ON MEGLASS. COMPARISON OF THE FACE VERIFICATION ACCURACY ON $M1$, $M2$, AND $M2_r$ TO EVALUATE THE EFFECTIVENESS OF EYEGLASSES REMOVAL. $M1$ CONTAINS FACE IMAGES WITHOUT GLASSES. $M2$ ONLY INCLUDES FACE IMAGES WITH GLASSES. $M2_r$ CONTAINS THE SAME IMAGES WITH $M2$ WHILE THOSE IMAGES ARE MANIPULATED WITH GLASSES REMOVAL ADOPTING OUR METHOD

Face images	Accuracy
Images without eyeglasses $M1$	96.34%
Images with eyeglasses $M2$	94.00%
Images after eyeglasses removal $M2_r$	94.66%

We show the qualitative face verification results on LFW in Fig. 10(a). We observe that the Euclidean distance decreases between matched images and the generated image by our method. Moreover, we present results of the average Euclidean distance in Fig. 10(b). The blue bar denotes the average Euclidean distance of the same person with and without eyeglasses, and the orange bar denotes the average Euclidean distance of the same person without glasses and removing glasses by ERGAN. We observe that on both LFW and MeGlass, the average Euclidean distance decrease between the same person without glasses and glasses removed by ERGAN. The result indicates that the proposed method could effectively alleviate the interference of glasses occlusion on face verification while preserving identification information well. Quantitative and qualitative results demonstrate that the proposed method is beneficial to the face verification task.

Facial Expression Recognition: Besides, we demonstrate that the proposed method benefits the facial expression recognition. To evaluate our method, we adopt a facial expression recognition algorithm,[2] which effectively classifies the emotion into six adjectives (i.e., angry, scared, happy, sad, surprised, and neutral). We perform experiments on the

[1]https://github.com/davidsandberg/facenet

[2]https://github.com/opconty/keras-shufflenetV2

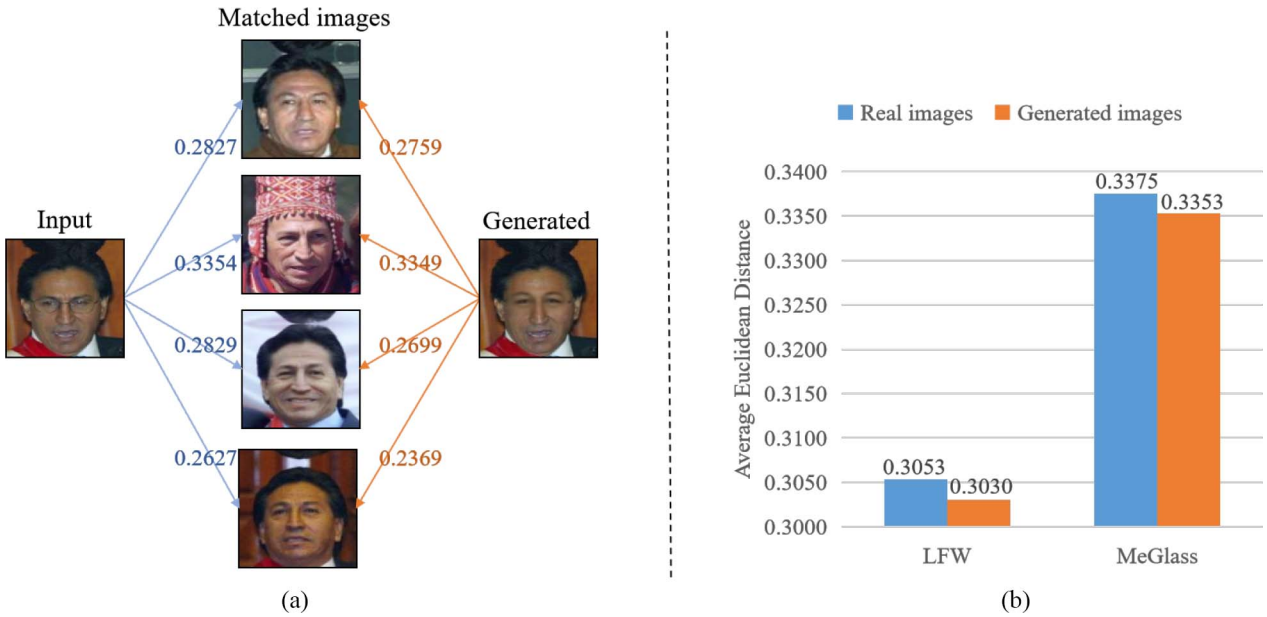


Fig. 10. Comparison of the Euclidean distance (lower is better) to evaluate the effectiveness of eyeglasses removal by our method. (a) Left image is one input face image with eyeglasses. The middle images are other face images of the same identity. The right image is the generated image which is removed glasses by our method. The number on the line denotes the Euclidean distance between two images. (b) Blue bar denotes the average Euclidean distance of the same person with and without eyeglasses, and the orange bar denotes the average Euclidean distance of the same person without glasses and removing glasses by ERGAN.

Fig. 11. Qualitative facial expression recognition results. The red word in the upper left corner of each image indicates the result of facial expression recognition. First row: expression recognition results of original images. Second row: expression recognition results of generated images by glasses removal.

TABLE VIII
QUANTITATIVE FACIAL EXPRESSION RECOGNITION RESULTS. COMPARISON OF THE FACIAL EXPRESSION RECOGNITION ACCURACY ON $S1$, $S2$, AND $S2_r$ TO EVALUATE THE EFFECTIVENESS OF EYEGLASSES REMOVAL. $S1$ CONTAINS FACE IMAGES WITHOUT GLASSES. $S2$ ONLY INCLUDES FACE IMAGES WITH GLASSES. $S2_r$ CONTAINS THE SAME IMAGES WITH $S2$ WHILE THOSE IMAGES ARE MANIPULATED WITH GLASSES REMOVAL ADOPTING OUR METHOD

Face images	Accuracy
Images without eyeglasses $S1$	80.61%
Images with eyeglasses $S2$	73.65%
Images after eyeglasses removal $S2_r$	78.59%

CelebA dataset. Notably, the CelebA dataset only labels the attribute of smiling or not. Accordingly, we view that smiling represents the emotion of happy, and no-smiling represents other emotions. During the test, three sets are taken into consideration. $S1$ denotes that the set contains face images with attributes of smiling and no-glasses, and $S2$ indicates that the set contains face images with attributes of smiling and glasses. $S2_r$ includes the same samples as $S2$, while all images are manipulated with eyeglasses removal by our method.

We show qualitative results of facial expression recognition in Fig. 11. We observe that face images with glasses in the first row are all misidentified as other expressions. In contrast, after the manipulation of eyeglasses removal by our method, all images are recognized as the happy expression in the second row.

We present quantitative results in Table VIII. Comparing the facial expression recognition results between $S1$ and $S2$, we find that the accuracy drops by 6.9%. This result indicates that facial expression recognition is affected by the occlusion of glasses. Besides, comparing the results between $S2$ and $S2_r$, the accuracy of expression recognition gains by 4.9%. The performance is close to the result on $S1$, which demonstrates the benefit of eyeglasses removal for facial expression recognition.

V. CONCLUSION

In this article, we have proposed a GAN-based framework for eyeglasses removal in the wild. We adopt a dual learning scheme simultaneously to learn two inverse manipulations (removing glasses and wearing glasses), which enforces the model to generate high-quality results. Extensive qualitative and quantitative experiments demonstrate that our method outperforms previous state-of-the-art methods in terms of realism and diversity. Furthermore, we remark that the proposed method has the potential to serve as a preprocessing tool for other face-related tasks, e.g., face verification and facial expression recognition.

REFERENCES

[1] L. Zebrowitz, *Reading Faces: Window to the Soul?* London, U.K.: Routledge, 2018.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 815–823.

[3] N. McLaughlin, J. Ming, and D. Crookes, "Largest matching areas for illumination and occlusion robust face recognition," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 796–808, Mar. 2017.

[4] Y. Guo, L. Jiao, S. Wang, S. Wang, and F. Liu, "Fuzzy sparse autoencoder framework for single image per person face recognition," *IEEE Trans. Cybern.*, vol. 48, no. 8, pp. 2402–2415, Aug. 2018.

[5] J. Gaston, J. Ming, and D. Crookes, "Matching larger image areas for unconstrained face identification," *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 3191–3202, Aug. 2018.

[6] Y. Wang, Y. Y. Tang, L. Li, and H. Chen, "Modal regression-based atomic representation for robust face recognition and reconstruction," *IEEE Trans. Cybern.*, early access, Mar. 21, 2019, doi: 10.1109/TCYB.2019.2903205.

[7] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, "Learning meta face recognition in unseen domains," 2020. [Online]. Available: arXiv:2003.07733.

[8] D. Cao, X. Zhu, X. Huang, J. Guo, and Z. Lei, "Domain balancing: Face recognition on long-tailed domains," 2020. [Online]. Available: arXiv:2003.13791.

[9] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1805–1812.

[10] P. Liu, J. T. Zhou, I. W.-H. Tsang, Z. Meng, S. Han, and Y. Tong, "Feature disentangling machine—A novel approach of feature selection and disentangling in facial expression analysis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 151–166.

[11] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 2168–2177.

[12] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017. [Online]. Available: arXiv:1708.04896.

[13] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2019.

[14] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 501–518.

[15] Y. Sun *et al.*, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 393–402.

[16] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 542–551.

[17] C. Wu, C. Liu, H.-Y. Shum, Y.-Q. Xu, and Z. Zhang, "Automatic eyeglasses removal from face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 322–336, Mar. 2004.

[18] J.-S. Park, Y. H. Oh, S. C. Ahn, and S.-W. Lee, "Glasses removal from facial image using recursive error compensation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 805–811, May 2005.

[19] C. Du and G. Su, "Eyeglasses removal from facial images," *Pattern Recognit. Lett.*, vol. 26, no. 14, pp. 2215–2220, 2005.

[20] D. Yi and S. Z. Li, "Learning sparse feature for eyeglasses problem in face recognition," in *Proc. IEEE Face Gesture*, 2011, pp. 430–435.

[21] W. K. Wong and H. Zhao, "Eyeglasses removal of thermal image based on visible information," *Inf. Fusion*, vol. 14, no. 2, pp. 163–176, 2013.

[22] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London Edinburgh Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.

[23] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2868–2881, Jun. 2017.

[24] X.-S. Wei, C.-L. Zhang, J. Wu, C. Shen, and Z.-H. Zhou, "Unsupervised object discovery and co-localization by deep descriptor transformation," *Pattern Recognit.*, vol. 88, pp. 113–126, Apr. 2019.

[25] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3730–3738.

[26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. NeurIPS*, 2017, pp. 6626–6637.

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 586–595.

[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 586–595.

[29] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. NeurIPS*, 2017, pp. 700–708.

[30] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 179–196.

[31] M. De Smet, R. Fransens, and L. Van Gool, "A generalized EM approach for 3D model based face recognition under occlusions," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, 2006, pp. 1423–1430.

[32] S. Mika *et al.*, "Kernel PCA and de-noising in feature spaces," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 536–542.

[33] I. J. Goodfellow *et al.* "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.

[34] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "CamStyle: A novel data augmentation method for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, Mar. 2019.

[35] Z. Yang, J. Dong, P. Liu, Y. Yang, and S. Yan, "Very long natural scenery image prediction by outpainting," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 10560–10569.

[36] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6777–6786.

[37] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5802–5810.

[38] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Adversarial style mining for one-shot unsupervised domain adaptation," 2020. [Online]. Available: arXiv:2004.06042.

[39] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1225–1233.

[40] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 422–437.

[41] M. Liu *et al.*, "STGAN: A unified selective transfer network for arbitrary image attribute editing," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3673–3682.

[42] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013. [Online]. Available: arXiv:1312.6114.

[43] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," in *Proc. NeurIPS*, 1994, pp. 3–10.

[44] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *Int. J. Comput. Vis.*, vol. 124, no. 3, pp. 409–421, 2017.

[45] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2868–2876.

[46] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.*, 2000, pp. 417–424.

[47] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2536–2544.

[48] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 4470–4479.

[49] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Rep. 07-49, 2007.

[50] G. B. Huang, M. A. Mattar, H. Lee, and E. G. Learned-Miller, "Learning to align from scratch," in *Proc. NeurIPS*, 2012, pp. 773–781.

[51] J. Guo, X. Zhu, Z. Lei, and S. Z. Li, "Face synthesis for eyeglass-robust face recognition," in *Proc. Chin. Conf. Biometr. Recognit.*, 2018, pp. 275–284.

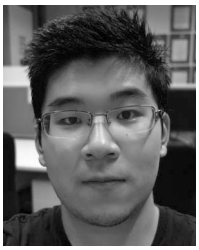
[52] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 4873–4882.

[53] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a GAN to learn how to do image degradation first," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 187–202.

[54] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2138–2147.
[55] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," 2019. [Online]. Available: arXiv:1903.05628.
[56] A. Razavi, A. V. D. Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," 2019. [Online]. Available: arXiv:1906.00446.
[57] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9.
[58] J.-Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 465–476.
[59] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 36–52.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
[61] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4105–4113.
[62] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1510–1519.
[63] 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 *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8798–8807.
[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: arXiv:1412.6980.
[65] X. Mao, Q. Li, H. Xie, R. Lau, Z. Wang, and S. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2813–2821.
[66] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?" in *Proc. ICML*, 2018, pp. 3478–3487.

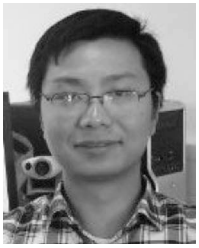
Bingwen Hu is currently pursuing the Ph.D. degree with the Nanjing University of Science and Technology, Nanjing, China.

He is also a joint Ph.D. student with the University of Technology Sydney, Ultimo, NSW, Australia. His research interests include computer vision, image processing, and pattern recognition.



Zhedong Zheng received the B.S. degree in computer science from Fudan University, Shanghai, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Computer Science, University of Technology Sydney, Ultimo, NSW, Australia.

His research interests include robust learning for image retrieval, generative learning for data augmentation, and unsupervised domain adaptation.

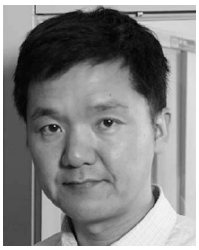


Ping Liu (Member, IEEE) received the bachelor's degree in electrical engineering from the Wuhan University of Technology, Wuhan, China, in 2005, the master's degree from the Huazhong University of Science and Technology, Wuhan, in 2008, and the Ph.D. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA, in 2015.

He is a Research Staff with the Center for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW, Australia. His research interests include computer vision, machine learning, and deep learning.

Wankou Yang received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Nanjing University of Science and Technology, Nanjing, China, in 2002, 2004, and 2009, respectively.

From 2009 to 2011, he was a Postdoctoral Fellow with the School of Automation, Southeast University, Dhaka, Bangladesh. From 2011 to 2016, he was an Assistant Professor with the School of Automation, Southeast University, where he is currently an Associate Professor with the School of Automation. His research interests include pattern recognition and computer vision.



Mingwu Ren received the Ph.D. degree in pattern recognition and intelligent systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2001.

He is currently a Professor with the School of Computer Science and Engineering, NUST. His research interests include computer vision, image processing, and pattern recognition.