IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 August 2023, accepted 12 August 2023, date of publication 21 August 2023, date of current version 25 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3307132

RESEARCH ARTICLE

Adv-Eye: A Transfer-Based Natural Eye Makeup Attack on Face Recognition

JIATIAN PI[1], JUNYI ZENG[2], QUAN LU[3], NING JIANG[3], HAIYING WU[3], LINCHENGXI ZENG[3], AND ZHIYOU WU[2]

[1]National Center for Applied Mathematics in Chongqing, Chongqing 401331, China
[2]Department of Mathematical Sciences, Chongqing Normal University, Chongqing 401331, China
[3]Mashang Consumer Finance Company Ltd., Chongging 400000, China

Corresponding author: Ning Jiang (ning.jiang02@msxf.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 11871128, and in part by the Chongqing Education Commission Key Project under Grant KJZD-K202114802.

ABSTRACT Deep face recognition models are vulnerable to adversarial samples generated by adversarial attack methods. However, current attack methods do not adequately represent the security problems of the deep FR models, because they either produce adversarial samples which are unnatural and easily perceived by human or have poor attack capabilities with low attack success rates on the black-box victim FR model. To achieve a good trade-off between the imperceptibility and attack capability, we propose Adv-Eye, a novel method for constructing adversarial facial images by adding natural eyeshadow to the orbital region. Adv-Eye consists of Makeup Generation Module, Makeup Blending Module, and Attack Module. Makeup Generation Module develops pre-makeup strategy to help adversarial generative networks (GANs) to accurately generate eyeshadow on the orbital image. Makeup Blending Module develops a multi-view image visual similarity evaluation method to improve the imperceptibility of the generated eyeshadow. In Attack Module, an ensemble attack strategy based on fine-grained meta-learning and input decay, is applied to improve attack capability under query-free black-box setting. From the experimental results under LADN dataset and MT dataset, compared with existing techniques, the adversarial samples generated by Adv-Eye not only significantly improve the visual quality, but also achieve average success rates of **1.63%** and **1.05%** improvement on the local black-box FR model and average confidence point improvements of **5.33** and **5.22** on the online commercial FR platform respectively. The above results demonstrate that pre-makeup strategy and multi-view image visual similarity evaluation method effectively improve the imperceptibility of generated adversarial perturbations, and Attack Module effectively improves attack success rate while maintaining high image quality.

INDEX TERMS Adversarial attack, generative adversarial networks, meta-learning, face recognition.

I. INTRODUCTION

With the advancement in Deep Neural Networks (DNNs), DNN-based approaches have achieved state-of-the-art and even surpassed human performance on various computer vision tasks [1], [2], [3], [4], [5]. As an important task, face recognition is also greatly advanced by DNNs [1], [6], [7]. Because of its excellent capability, deep face recognition models are used in many security-sensitive authentication scenarios (*e.g.* face-sweeping payment, smart device unlocking, and public access). However, recent works have shown that DNNs are vulnerable to adversarial samples which are generated by adding elaborately designed imperceptible perturbations to clean images [8], [9], [10]. In face recognition, inevitably, adversarial samples also exist and effectively work on the state-of-the-art face recognition (FR) systems which have been widely applied in real-world scenarios [11], [12]. These findings have raised significant

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

concerns regarding the security and reliability of current DNN-based FR systems.

However, current adversarial attack methods on FR systems have inherent limitations, which means they do not fully represent the security risks of deep face recognition systems in practical application. Specifically, the major limitations are the following: (1) **Low attack capability:** most existing methods belong to white-box attack (*i.e.* having full knowledge of the victim models especially gradient information) [13], [14], [15] or query-based black-box attack (*i.e.* attacker is able to arbitrarily query the victim model) [11], [12]. Without any information of the victim model, they hardly attack successfully. Therefore, they are hardly effective in practical applications. (2) **Lack of naturalness:** the adversarial perturbations generated by current methods lack of naturalness and realism. Specifically, several methods set the form of adversarial perturbations to noise [11], [12], [16]. This form often results in the modified facial image appearing unnatural, making it easily noticeable to the human eye, and using image preprocessing techniques (*e.g.* face alignment, face cropping and image compression) to disrupt the perturbation structure, leading to attack failure [17]. Some adversarial attack methods try to solve it by imbuing adversarial perturbations with specific semantic meanings, Brown et al. [13] set perturbations to square patch, Komkov and Petiushko [15] bend the patch by an affine transformation and set it on the front of the hat. Whereas, the disorderly arrangement of pixel values within adversarial patches and the abrupt change of pixel values at the boundary between adversarial patches and source images make generated adversarial samples have obvious image manipulation traces. In order to alleviate manipulation traces, some attack methods combine GANs to generate adversarial perturbations which better fit the human face. Jia et al. [18] generate adversarial samples by editing face attributes (*e.g.* adding smile and adding wrinkles). Lin et al. [19] generate adversarial samples by incorporating full-face makeup onto facial regions. Hu et al. [20] incorporate a regularization module into the generative model to enhance the visual realism of adversarial makeup. However, such global modifications can easily lead to unexpected changes in sensitive identity attributes, thereby compromising the identity consistency between adversarial facial images and source images in human visual perception. Yin et al. [21] propose Adv-Makeup framework, which attempt to combine the ideas of the above two types of semantic perturbations to improve visual quality. Adv-Makeup incorporates the idea of using eyeshadow as the form of adversarial perturbations. Adv-Makeup utilizes GANs to generate the eye-region images with eyeshadow and merge them to the source images. Furthermore, Adv-Makeup introduces an integrated attack approach based on fine-grained meta-learning to enhance the attack capability of the generated adversarial samples. Nevertheless, in Figure 1 (b), the generated makeup adversarial samples still appear too conspicuous and exhibit unexpected changes in semantic content. Moreover, there is still room for improvement in the attack success rate of the generated adversarial samples.

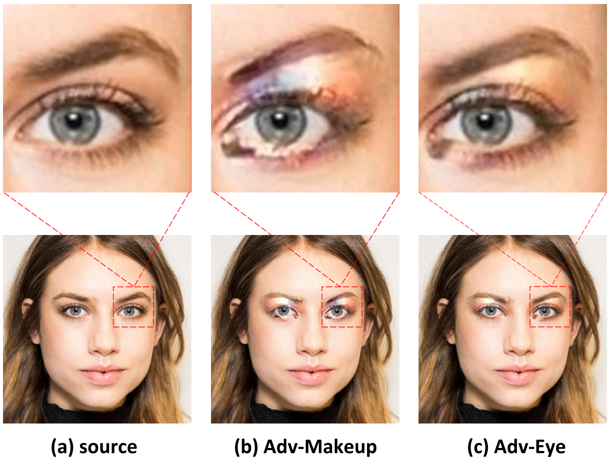


FIGURE 1. **Illustration of Adv-Makeup and Adv-Eye. (a) Source facial image and the corresponding zoomed-in image. (b) Adversarial facial image and the corresponding zoomed-in image generated by Adv-Makeup. (c) Adversarial facial image and the corresponding zoomed-in image generated by Adv-Eye.**

In conclusion, the main challenge of effectively misleading deep face recognition models under real-world conditions lies in simultaneously achieving excellent imperceptibility of adversarial samples and a high attack success rate under query-free black-box setting (*i.e.* the transferability of attack). To face this challenge, we propose a novel attack method called Adv-Eye, which aims to generate adversarial samples with both high visual quality and high attack success rate. Similar to Adv-Makeup, Adv-Eye also utilizes eyeshadow as the form of adversarial perturbations. For improving the quality of generated images, we introduce the pre-makeup method to assist in the training of GANs and design a multi-view image consistency metric to simulate human vision and quantify the naturalness of generated images. The resulting synthesized faces with eyeshadow, as shown in Figure 1 (c), appear more natural to human eyes. Additionally, in order to further enhance the attack effect without relying on the victim model, we propose an ensemble attack method which combines input decay with fine-grained meta-learning to further improve the transferability of attack while maintaining its natural appearance. Our contributions can be summarized as follows:

- We propose Adv-Eye, a novel approach that can simultaneously attain high imperceptibility and attack transferability against face recognition models by adding eyeshadow.
- We develop a pre-makeup method based on Poisson fusion, which assists GANs in synthesizing eyeshadow more accurately on the orbital region of facial images.
- To enhance the consistency of the generated eyeshadow with the source image in terms of human vision and improve the imperceptibility of the attack behavior, we incorporate two image quality metrics SSIM and

Lpips to measure the naturalness of images from different perspectives and quantify the visual quality of the images more effectively.

- We propose an ensemble attack strategy based on fine-grained meta-learning and input decay. During meta-train phase, we integrate meta-train models and performs multi-step training. During meta-test phase, we construct meta-test models by introducing input decay. This allows us to better utilize the information of white-box surrogate models while improving the transferability of adversarial samples to black-box victim models.
- The comprehensively designed experiments on LADN dataset and makeup transfer (MT) dataset demonstrate that the adversarial samples generated by Adv-Eye exhibit both better image quality and attack performance against offline face recognition models and online commercial platforms.

II. RELATED WORKS

In this section, we provide a brief overview of the related works in the area of Adversarial Attack, Adversarial Attack on Face Recognition, Generative Adversarial Networks, and Meta-Learning.

A. ADVERSARIAL ATTACK

The core of Adversarial Attack is quite simple: let us slightly modify the image input to the neural network model so that the classification result output by the network changes from the correct class to another. Based on this idea, szegedy et al. [8] formulate the adversarial attack problem as the following optimization model:

$$\begin{aligned}&\min_{r}\|r\|_2\\ &\mathtt{s.t.}\ f(x)=y_{true}\\ &\qquad f(x+r)=y_t\neq y_{true}\\ &\qquad x+r\in[0,1]^m\end{aligned}\tag{1}$$

where r represents the perturbation, $x \in [0,1]^m$ is the input image of size m, $f(x)$ represents the classification result of model f corresponding to x, y_t is the target label, and y_{true} is the ground-truth label. Note that in Eq. (1), if attackers simply want $f(x+r)\neq y_t$, the attack is called untargeted attack (dodging attack in case of face recognition), and if attackers need a specific predefined class y_t, the attack is called targeted attack (impersonation attack in case of face recognition). Besides, Szegedy et al. [8] propose a quasi-Newton L-BFGS-B method to solve the above optimization problem. In [9], Goodfellow et al. consider gradient backpropagation mechanism of network training, and propose a simpler but more effective method named Fast Gradient-Sign Method (FGSM). FGSM obtains the adversarial sample by using the gradient information of the victim model as follows:

$$x_{adv}=\begin{cases}x+\alpha\mathrm{sign}(\nabla_x J(f(x),y_{true})) & \text{untargeted}\\ x-\alpha\mathrm{sign}(\nabla_x J(f(x),y_t)) & \text{targeted}\end{cases}\tag{2}$$

where x_{adv} represents the resulting adversarial sample, J is loss function which depends on the task of the victim model (*e.g.* cross-entropy function used for classify task). α is the hyperparameter that controls the size of the perturbation. To further increase the attack success rate and reduce the perturbation which is required for successful attack, Kurakin et al. [22] propose basic iterative Method (BIM). In BIM, the single-step attack method is changed to a small step-length multiple attack method. Furthermore, Madry et al. [23] propose project gradient descent (PGD) method. Compared with BIM, PGD changes the initial perturbations from 0 to random perturbations and controls the size of perturbations using the maximum inner product projection method.

However, the above approaches require all the information of the victim model, this is not in line with practical scenario. Some researches have revealed that adversarial samples generated against one model may mislead other models under the same task [8], [24], [25]. This phenomenon is known as the transferability of adversarial attack. Moreover, the transferability of adversarial attack can be analogous to the generalization of a trained model. Therefore, referring to the common tricks for improving model generalization, a series of methods have been proposed to improve the transferability from three perspectives: input transformation [26], [27], [28], [29], optimal gradient orientation [28], [30], [31], and model ensemble [31], [32]. In our work, we design an ensemble attack strategy based on fine-grained meta-learning and input decay to achieve more powerful black-box attack.

B. ADVERSARIAL ATTACK ON FACE RECOGNITION

Although adversarial attack method only targets image classification task at first, a series of studies have indicated that adversarial samples are also present in other vision tasks [33], [34], [35]. Face recognition is one of the key areas, therefore, many studies have been proposed for face recognition to examine the robustness of recent FR models. Deb et al. [16] generate face adversarial samples through adversarial generative networks (GANs). Cherepanova et al. [35] propose LowKey method, which introduces a perceptual loss function based on LPIPS [36] to reduce the gap between the generated adversarial samples and the source image. Yang et al. [12] propose TIP-IM method which utilizes maximum mean discrepancy (MMD) instead of L_p parametric to measure the naturalness of the adversarial sample. Moreover, if the target identity corresponds to multiple face images, TIP-IM introduces an optimal strategy. In each iteration, the image that is currently the most suitable as the target is selected to improve the attack effectiveness.

Above methods generate adversarial samples by adding noise perturbations to faces. However, as shown in the Figure 2(a), such methods create awful distortion on facial images which cause the adversarial sample to be easily detected. Thus, as shown Figure 2(b), a series of studies make generated perturbations semantically meaningful to improve

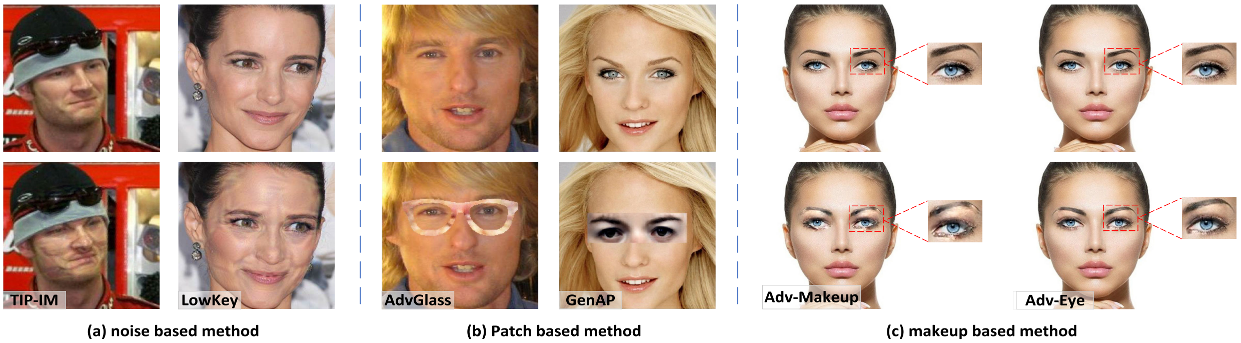


FIGURE 2. Comparison with existing adversarial attack methods against FR models. (a) Noise-based attack methods: TIP-IM [12], LowKey [35]. (b) Patch-based methods: AdvGlass [14], GenAP [37]. (c) Makeup-based methods: Adv-Makeup [21], Adv-Eye in this paper.

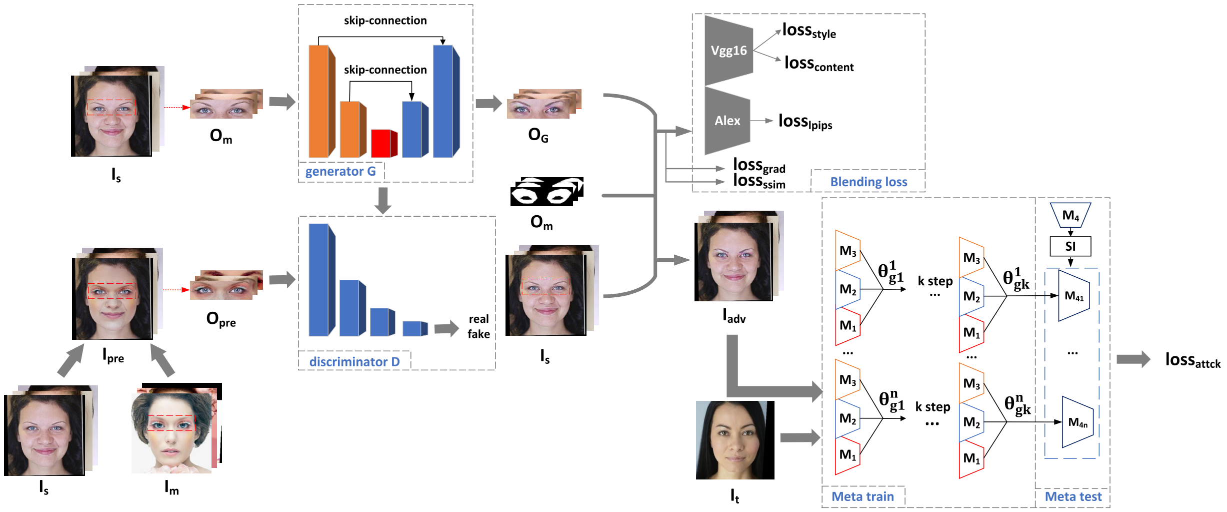
naturalness and imperceptibility of the adversarial sample. Komkov and Petiushko [15] propose Adv-Hat model, which sets adversarial perturbations to stickers on the front of hats. Sharif et al. [14], [38] propose Adv-Glasses model, which implants adversarial perturbations into eyeglasses to mislead the face recognition system. Xiao et al. [37] propose Gen-AP model, which generates adversarial stickers with patterns of eyes based on a pre-trained StyleGAN [39] model. Whereas, the perturbations generated by the above methods are rather conspicuous and not natural enough. Facing this problem, Zhu et al. [40] firstly attempt using makeup as the form of adversarial perturbations, and generating adversarial eye makeup on source images to mislead face recognition (FR) systems. In order to remove the white-box limitation in [40] and further improve the imperceptibility of the generated eyeshadow, Yin et al. [21] propose Adv-Makeup model which generates the eye makeup adversarial images by GANs. In terms of improving imperceptibility, Yin et al. introduce a blending method based on image edge consistency and high-dimensional feature consistency. To achieve attack in black-box setting, Yin et al. propose an ensemble attack method based on fine-grained meta-learning to implement transferable black-box attack. Whereas, Adv-Makeup has the following limitations: (1) As shown in left part of Figure 2(c), visual artifacts exist in the makeup area. (2) The generated adversarial samples lack transferability. In this paper, for addressing the aforementioned limitations, we propose the Adv-Eye model, which generates more natural eyeshadow while enhancing the attack transferability of the generated adversarial samples.

C. GENERATIVE ADVERSARIAL NETWORKS

Goodfellow et al. [41] firstly propose GANs to generate handwritten digital images through noise vectors. Adversarial Generative Networks consist of two parts: the generator G, which generates images based on input, and the discriminator D, which distinguishes between generated and reference images. By alternatively training G and D, the generated image distribution converges towards the reference image distribution, leading to enhanced naturalness of the generated images. Isola et al. [42] propose pix2pix approach, which introduces U-Net structure in G and incorporates the source image into the input of D to achieve image-to-image translation. Zhu et al. [43] propose CycleGAN, which design a cycle consistency loss to relax the paired data requirement of pix2pix. However, CycleGAN requires reference images to have consistent texture properties (such as the same color or pose). Due to the diversity of eyeshadows, the images generated by CycleGAN exhibit noticeable visual artifacts. To address this issue, Adv-Makeup improves the naturalness of the generated images by using losses based on edge consistency and high-dimensional feature consistency. However, the images generated by Adv-Makeup still exhibit significant visual artifacts and undesirable changes in semantic content. In this paper, to further maintain semantic consistency, we introduce pre-makeup method to enhance the semantic content consistency between the reference and source images, thereby better preserving the semantic consistency of the generated images. Moreover, besides edge consistency and high-dimensional feature consistency, we incorporate SSIM and Lpips to measure the visual quality of generated images based on pixel statistics features and perceptual distance. This approach comprehensively assesses image naturalness and reduces the generation of visual artifacts.

D. META LEARNING

The goal of meta-learning is to enable models to perform well on different tasks and the core idea of meta-learning is ‘‘learn to learn’’ [44]. Different from conventional training processes, Meta-learning methods learn the connections between different tasks through the processes of meta-train and meta-test [45], [46]. Yuan et al. [47] divide the surrogate image classification models into meta-train models and meta-test models to find the common attack direction of the surrogate models by meta-train and meta-test, thereby enhancing the transferability of adversarial samples. In the Adv-Makeup framework, Yin et al. [21] propose an ensemble strategy based on fine-grained meta-learning which combines the gradient information from meta-train step with the final gradient to stabilize gradient update direction. However, Adv-Makeup individually performs single-step meta-train for each meta-train model and repeatedly uses the same meta-test



(a) Makeup Generation Module (b) Makeup Blending Module (c) Attack Module

FIGURE 3. Overview of Adv-Eye framework, which consists of three modules: (a) Makeup Generation Module based on GANs and pre-makeup method. (b) Makeup Blending Module based on multi-perspective image consistency evaluation strategy. (c) Attack Module based on Fine-grained meta-learning and input decay.

model, which still does not fully leverage the information of the surrogate models, resulting in insufficient transferability of the adversarial samples. In order to more fully utilize the information of the surrogate models, we propose an ensemble attack method based on meta-learning and input decay. We integrate meta-train models and conduct multi-step training in meta-train phase. During meta-test phase, we add different decay layers to the meta-test model.

III. METHODOLOGY

A. OVERVIEW

This section describes the optimization problem abstracted from the Adv-Eye model and introduces each of the modules that constitutes Adv-Eye. In general, Adv-Eye model can be summarized by the following optimization model:

$$\min_{\theta_{gen}}[\max_{\theta_{dis}} \mathrm{V}(P_G, P_{data}) + \lambda_I \mathrm{IM}(I_{adv}, I_s) - \lambda_{at} \mathrm{J}(I_{adv}, I_t)]. \tag{3}$$

where θ_{gen} are parameters of generator, θ_{dis} are parameters of discriminator. $\mathrm{V}(P_G, P_{data})$ denotes a distance metric of generated image distribution versus reference image distribution. $\mathrm{IM}(I_{adv}, I_s)$ denotes the naturalness metric of eye makeup adversarial sample. $\mathrm{J}(I_{adv}, I_t)$ represents attack transferability of the eye makeup adversarial sample. The framework of Adv-Eye is shown in Figure 3 and consists of three main modules: Makeup Generation Module, Makeup Blending Module, and Attack Module. Makeup Generation Module generates facial images with eyeshadow by inputting a non-makeup source facial image and a makeup reference image. Makeup Blending Module helps to improve the naturalness and visually-indistinguishable quality of the generated images. To achieve powerful black-box attack, Attack Module, which includes an ensemble attack strategy based on fine-grained meta-learning and input decay, is introduced to improve the transferability of generated adversarial samples.

B. MAKEUP GENERATION

The aim of Adv-Eye is to fool FR systems by adding natural-looking adversarial makeup to the source human face. Considering the widespread use of eyeshadow in face makeup and the eye area being one of the most important discriminative areas for face recognition models, the form of the adversarial perturbation generated by Adv-Eye is designed as eyeshadow. Therefore, the core of Adv-Eye is finding a mapping that translates a source image into the image with no change other than adversarial eyeshadow. As shown in Figure 4, Makeup Generation Module contains a generator G as the mapping network to synthesize eyeshadow and a discriminator D to distinguish between real makeup images and generated images in order to improve perceptual realism of generated eyeshadow.

Specifically, $\mathrm{V}(P_G, P_{data})$ is transformed to the generator loss L_{gen} and the discriminator loss L_{dis}. G and D are trained alternatively based on L_{gen} and L_{dis} to find a Nash equilibrium solution to the min-max game in Eq. (3). However, obtaining paired makeup face dataset (*i.e.* plain faces and corresponding makeup faces taken under the same condition) is quite difficult. If training with unpaired data, D can hardly guide G in synthesis precisely, which leads to a low natrualness of generated images. Inspired by [48], Makeup Generation Module uses pre-makeup image I_{pre}, which generated by image warping and Poisson Fusion [49] as makeup reference image. Compared to directly using unpaired makeup reference image I_m, I_{pre} highlights the difference between source images and makeup reference images in eye makeup. Thus, D can guide G more effectively.

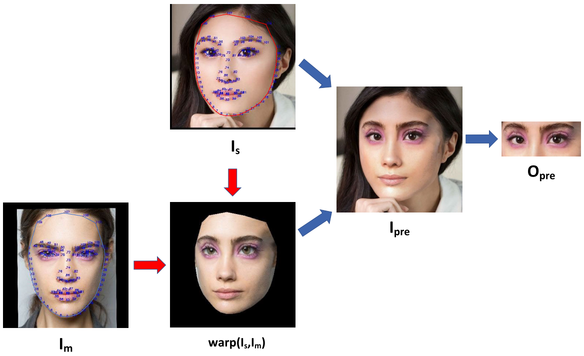


FIGURE 4. Illustration pre-makeup method. Red arrows indicate the process of warping the makeup reference image to be the same pose as the source image by landmarks. Blue arrows indicate the process of fusing the after-warping image and the source image.

Note that fusion results sometimes possess artifacts, which can be fixed by the network in generated results. The main process of pre-makeup is shown in Figure 4.

Specifically, with the input of source image I_s and real-world makeup image I_m. Firstly, we align I_s and I_m to extract their face landmarks, source image's orbital region $O_s \sim P_s$, and generate the corresponding binary mask M. Generator G takes O_s as input and outputs $O_G = G(O_s)$ with synthetic eyeshadow. Then, the resulting makeup orbital region O_F can be calculated as follows:

$$O_F = O_s \odot (1 - M) + O_G \odot M, \tag{4}$$

where $\odot$ means Hadamard product. Meanwhile, O_G is input to D and calculates loss_{gen} for the training of G. Finally, O_G is attached to the orbital region of I_s to get resulting makeup face I_{adv}. For the training of discriminator D, we get pre-makeup image $I_{pre} = \text{W}(I_s, I_m)$ and corresponding orbital region O_{pre}, where W represents pre-makeup process. Then, D takes O_G and O_{pre} to calculate L_{dis}. The generator loss L_{gen} and the discriminator loss L_{dis} can be denoted as follows:

$$L_{gen} = \mathbb{E}_{O_s \sim P_s}\left[\log\left(1 - D\left(O_G\right)\right)\right]. \tag{5}$$

$$L_{dis} = -\mathbb{E}_{O_s \sim P_s}[\log\left(D\left(O_{pre}\right)\right) + \log\left(1 - D\left(O_G\right)\right)]. \tag{6}$$

C. MAKEUP BLENDING

Through Makeup Generation Module, generator G is able to generate images with eyeshadow. However, directly pasting the synthesized orbital region to the source image by using a binary mask produces obvious artifacts such as obvious differences between the pasted area in the generated image and the source image, and apparent changes at the boundary. To eliminate these noticeable artifacts, we propose Makeup Blending Module as IM in Eq. (3) to measure image naturalness in multiple views. Makeup Blending Module consists of the following members: L_{grad}, L_{style}, $L_{content}$, L_{ssim}, and L_{lpips}.

To alleviate changes at the edges of orbital region in I_{adv}, Makeup Blending Module introduces a gradient-based edge consistency constraint and translates it into a differentiable loss function L_{grad}, by minimizing the loss function, the edges of I_{adv} is closer to I_s. Gradient constraint loss L_{grad} is defined as:

$$L_{grad} = \|[\Delta I_s \odot (1 - M^*) + \Delta \text{h}(O_G) \odot M^*] - \Delta I_s]\|_2^2, \tag{7}$$

where Δ represents the image gradient operator, and M^* represents zero padding of M, which has the same size as I_s. $\text{h}(O_G)$ means zero padding of O_G, which has the same size as I_s.

To encourage G to generate the eyeshadow which matches the style of source image I_s, Makeup Blending Module introduces SSIM [50] to measure the similarity of two images based on low-dimensional features (*i.e.* luminance, contrast, and image structure). L_{ssim} based on SSIM is defined as follows:

$$L_{ssim} = 1 - \text{SSIM}(O_s \odot (1 - M) + O_G \odot M, O_s), \tag{8}$$

where SSIM(x, y) is composed of three terms:

$$\text{SSIM}(x, y) = [\mu(x, y)]^{\alpha}[l(x, y)]^{\beta}[s(x, y)]^{\gamma},$$

$$\mu(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1},$$
$$l(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2},$$
$$s(x, y) = \frac{2\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}, \tag{9}$$

$\mu(x, y)$ measures the difference in image luminance between x and y, where μ represents mean gray value of the image. $l(x, y)$ measures the difference in image contrast between x and y, where σ represents variance of the image. $s(x, y)$ measures the difference in image structure between x and y, where σ_{xy} represents the covariance of x and y. c_1, c_2, and c_3 are constants to avoid the denominator being 0. α, β, and γ are weight hyperparameters.

In addition to the low-dimensional features, it is necessary to measure the consistency of high-dimensional features. Inspired by [51] and [52], Makeup Blending Module extracts high-dimensional features by pre-trained vgg16 model and calculates style loss L_{style} and content loss $L_{content}$ as follows:

$$L_{\text{style}} = \sum_{l \in l_s} \alpha_l \text{MSE}(\text{gm}(\text{Vgg}_\text{l}(O_G)), \text{gm}(\text{Vgg}_\text{l}(O_s))), \tag{10}$$

$$L_{content} = \sum_{l \in l_c} \beta_l \text{MSE}(\text{A}(\text{Vgg}_\text{l}(O_G)), \text{A}(\text{Vgg}_\text{l}(O_s))), \tag{11}$$

where l_s, l_c represent middle layers of vgg network where the style information and content information are extracted respectively. Vgg_l represents the feature map output from middle layer l. $\text{A}(\cdot) \in R^{N_l \times M_l}$ represents the flattened matrix corresponding to the feature map. MSE$(\cdot)$ represents mean square error. $\mathrm{gm}(\cdot) = \mathrm{A}(\cdot)\mathrm{A}^{\mathrm{T}}(\cdot) \in R^{N_l \times N_l}$ is the Gram matrix of the feature map. N_l is the number of channels in the feature map, and M_l is the number of elements per channel in the flattened feature map. α_l and β_l are the weights to balance the contribution of each layer when calculating L_{style} and $L_{content}$.

Furthermore, Makeup Blending Module introduces LPIPS [36] to measure the perceptual distance between the resulting makeup image and the source image and calculate $L_{lpips} = \mathrm{LPIPS}(I_{adv}, I_s)$. Makeup Blending Module measures the similarity between the generated makeup face image and the source image from multiple perspectives by combining the above loss functions, and guides G to generate eyeshadow which can blend more naturally into the source image. Makeup Blending Module calculates blend loss L_{blend} as follows:

$$\begin{aligned} L_{blend} &= \lambda_{grad} L_{grad} + \lambda_{ssim} L_{ssim} + \lambda_{style} L_{style} \\ &\quad + \lambda_{content} L_{content} + \lambda_{lpips} L_{lpips} \end{aligned} \tag{12}$$

where λ_{grad}, λ_{ssim}, λ_{style}, $\lambda_{content}$, λ_{lpips} resprent weight parameters of L_{grad}, L_{ssim}, L_{style}, $L_{content}$ and L_{lpips}.

D. ATTCK MODULE

1) ADVERSARIAL LOSS ON FR SYSTEMS

The main process of deep face recognition model work involves extracting feature vectors of two face images by the network and calculating the similarity between the feature vectors (*e.g.* cosine similarity). If the similarity is less than a specific threshold, the two images are determined to be the same identity. Therefore, this paper uses cosine similarity as attack loss L against the FR model. Note that as impersonation attack is more difficult and practical than dodging attack, Adv-Eye primarily focuses on impersonation attack. Impersonation attack loss L can be expressed as follows:

$$L = 1 - \cos(\mathrm{M}(I_{adv}), \mathrm{M}(I_t)), \tag{13}$$

where I_t denotes the face image belonging to the target identity which is different from the identity of the source image. $\mathrm{M}(\cdot) \in R^n$ denotes the feature vector extracted by model M. By decreasing attack loss, generator G is encouraged to generate the eyeshadow which can make the feature vector of I_{adv} closer to that of the target image I_t. However, such adversarial samples have limited transferability because they can easily "overfit" to the surrogate model. While they may mislead the surrogate model, they are not effective in successfully attacking the black-box victim model. To improve the transferability of adversarial samples, recent studies have introduced ensemble training to obtain more generalized adversarial perturbations by attacking multiple models simultaneously [28], [32]. Although the perturbations generated by such ensemble attack methods are more generalized, the simple synthesis by loss or feature does not fully utilize the information of the surrogate models. Therefore, the transferability of the generated adversarial sample remains insufficient. To further improve the transferability of the makeup adversarial sample, we propose Attack Module as J in Eq. (3), which combines input transformation with fine-grained meta-learning to more fully use the information of surrogate models and improve the attack success rate on the black-box victim FR model. The details of Attack Module are shown in Figure 5.

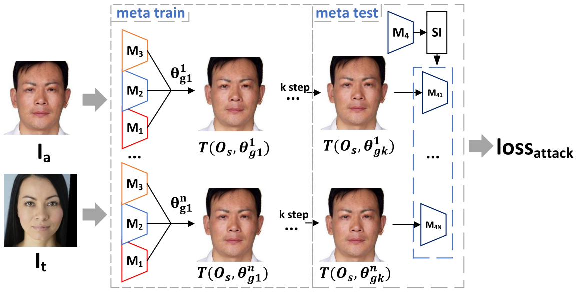


FIGURE 5. Illustration of the Attack Module structure. The above figure shows three meta-train models and a single base meta-test model as an example.

2) META-TRAIN

Attack Module randomly selects m pre-trained FR models from the FR model zoo Z, and use $m-1$ of them as meta-train models. The remaining one is used to build N meta-test models by adding different input decay layers.

In the meta-train step of nth meta-test model, Attack Module performs k-step iterations. In each iteration, the attack losses against all meta-train models are summed up as L^n_k. Then, the meta-gradient $\nabla^n_k \theta_g$ and intermediate generator parameters θ^n_{gk} are calculated as follows based on L^n_k:

$$\begin{aligned} L^n_k &= \frac{1}{m-1} \sum_{i=1}^{m-1} L(\mathrm{M_i}(I_t), \mathrm{M_i}(\mathrm{T}(O_s, \theta^n_{gk-1}))), \\ \nabla^n_k \theta_g &= \frac{\partial L^n_k}{\partial \theta_g}, \\ \theta^n_{gk} &= \theta^n_{gk-1} - \eta \nabla^n_k \theta_g, \end{aligned} \tag{14}$$

where $n \in \{1, \cdots, N\}, k \in \{1, \cdots, K\}$. $\mathrm{T}(O_s, \theta^n_{gk-1})$ denotes the makeup facial image generated by Makeup Generation Module using generator parameters θ^n_{gk-1}, η represents the learning rate in meta-train.

3) META-TEST

After the meta-train process, Attack Module conducts impersonation attack against every meta-test model M_{mn} using corresponding updated intermediate parameters θ^n_{gK}. After attacking meta-test models, the meta-test loss L^n_{te} is calculated as follows:

$$L^n_{te} = \mathrm{L}(\mathrm{M_{mn}}(I_t), \mathrm{M_{mn}}(\mathrm{T}(O_s, \theta^n_{gK}))). \tag{15}$$

Algorithm 1 *Adv-Eye* Attack Algorithm

```
Input: Source facial image $I_s$; makeup facial image $I_m$; target image $I_t$; pre-trained FR model zoo $Z$; initial generator parameters $\theta_g$; initial discriminator parameters $\theta_{dis}$; global iteration steps $T$; meta-train iteration steps $K$; number of meta-test models $N$; optimizer $Adam$; hyperparameters $\alpha_l$, $\beta_l$, $\alpha$, $\beta$, $\gamma$, $\eta$, $\lambda_{grad}$, $\lambda_{ssim}$, $\lambda_{style}$, $\lambda_{content}$, $\lambda_{lpips}$, $\lambda_{gen}$, $\lambda_{blend}$, $\lambda_{attack}$.
Output: Well-trained generator parameters $\theta_g^*$.
for $t = 0 \rightarrow T - 1$ do:
    Update discriminator $D$:
     Generate pre-makeup image W($I_s$, $I_m$);
    calculate $L_d is$ in Eq.(6);
     Update $\theta_{dis}$:
            $\theta_{dis} \leftarrow Adam(\theta_{dis}, L_{dis})$
    Update generator $G$:
     Calculate $L_{gen}$ in Eq. (5), $L_{blend}$ in Eq. (7)-(12);
    Meta-train:
     Random select $\mathrm{M_1}, \cdots \mathrm{M_m}$ pre-trained FR models from $Z$, $\mathrm{M_1}, \cdots \mathrm{M_{m-1}}$ as meta-train models. The rest FR model $\mathrm{M_m}$ is used to generate Meta-test model $\mathrm{M_{m1}}, \cdots \mathrm{M_{mN}}$.

    for $n = 1 \rightarrow N$ do:

        for $k = 1 \rightarrow K$ do:
            Calculate $L_k^n$, $\theta_{gk}^n$ in Eq.(14);
        end for
         Calculate $L_{te}^n$ in Eq. (15);
    end for
     Calculate $L_{attack}$ in Eq. (16);
     Calculate $L_G$ in Eq. (17);
     Update $\theta_g$:
            $\theta_g \leftarrow Adam(\theta_g, L_G)$
end for
return $\theta_g^* = \theta_g$.
```

After the processes of meta-train and meta-test, Attack Module combines the information obtained from the two stages to calculate the final attack loss L_{attack}, L_{attack} is calculated as follows:

$$L_{attack} = \sum_{i=1}^{N} (\frac{1}{N} L_1^n + L_{te}^n). \tag{16}$$

Ultimately, by combining the output of the three modules, the total loss of G is calculated as follows:

$$L_G = \lambda_{gen} L_{gen} + \lambda_{blend} L_{blend} + \lambda_{attack} L_{attack}. \tag{17}$$

where λ_{gen}, λ_{blend}, λ_{attack} represent balance weights. Overall, the entire training process of Adv-Eye is illustrated in Algorithm 1.

TABLE 1. The default parameter settings of Adv-Eye.

Name	Default	Description
α, β, γ	$(1, 1, 1)$	balance weights of u, l c in L_{ssim}
α_l	10^4	layer balance weight of L_{style}
β_l	1	layer balance weight of $L_{content}$
λ_{grad}	1	weight of L_{grad} in L_{blend}
λ_{style}	1	weight of L_{style} in L_{blend}
$\lambda_{content}$	1	weight of $L_{content}$ in L_{blend}
λ_{ssim}	1	weight of L_{ssim} in L_{blend}
λ_{lpips}	1	weight of L_{lpips} in L_{blend}
λ_{gen}	1	weight of L_{gen} in L_G
λ_{blend}	1	weight of L_{blend} in L_G
λ_{attack}	1	weight of L_{attack} in L_G
T	400	number of global iteration
M	4	number of white-box surrogate model used in Attack Module
N	3	number of meta-test FR model
K	5	number of meta-train iteration
η	10^{-3}	learning rate of meta-training
lr of generator training	10^{-3}	learning rate of generator training
$beta_1$ of generator optimizer	0.5	momentum weight of generator's Adam optimizer
$beta_2$ of generator optimizer	0.999	momentum weight of generator's Adam optimizer
weight decay rate of generator optimizer	10^{-4}	weight decay rate of generator's Adam optimizer
lr of discriminator training	4×10^{-4}	learning rate of discriminator training
$beta_1$ of generator optimizer	0.5	momentum weight of discriminator's Adam optimizer
$beta_2$ of generator optimizer	0.999	momentum weight of discriminator's Adam optimizer
weight decay rate of generator optimizer	10^{-4}	weight decay rate of discriminator's Adam optimizer

IV. EXPERIMENTS

A. EXPERIMENTAL SETTING

1) IMPLEMENTATION DETAILS

The structure of G and D follows LADN [53]. The optimizers for training G and D are respectively set to Adam optimizer [54]. The hyperparameter settings and meanings of Adv-Eye are shown in Table 1. Furthermore, the results of all our experiments are conducted on Titan XP 12GB*1.

2) COMPETITER

To verify that the adversarial samples generated by our Adv-Eye possess outstanding image quality and imperceptibility while maintaining high attack transferability on black-box models. We compare our approach with several benchmark adversarial attack methods. Specifically, the benchmark methods include BIM [55], DI-SI-CI-MI-FGSM [28], [30], [32], and Adv-Makeup [21]. BIM is a classic gradient-based iterative adversarial attack method. DI-SI-CI-MI-FGSM combines multiple input transformation methods (image resizing and zero-padding, cutout, input pixel value decay) and momentum gradient descent method to achieve a strong black-box attack ability. Adv-Makeup is the latest proposed physical adversarial attack method for deep FR models. Adv-Makeup also defines adversarial perturbation in the form of eye makeup, aiming to add eye makeup to source face images to generate adversarial face images and achieve black-box transferable impersonation attack.

In terms of hyperparameter settings for each competitor, the hyperparameter settings and meanings of BIM, DI-SI-CI-MI-FGSM, and Adv-Makeup are shown in Table 2, Table 3, and Table 4 respectively. It is worth noting that the parameter settings of BIM and DI-SI-CI-MI-FGSM are consistent with [56] and the parameter settings of Adv-Makeup are consistent with [21].

3) DATASETS

Two widely used public makeup datasets are used to test the effectiveness of each method. (1) LADN dataset [53], which includes 333 high quality frontal before-makeup faces and

TABLE 2. The default parameter settings of BIM.

Name	Default	Description
ε	20	constraint of the adversarial pertubations under L_∞
T	50	number of iteration
lr	0.8	weight of L_{grad} in L_{blend}

TABLE 3. The default parameter settings of DI-SI-CI-MI-FGSM.

Name	Default	Description
ε	20	constraint of the adversarial pertubations under L_∞
T	50	number of iteration
lr	0.8	weight of L_{grad} in L_{blend}
P_{DI}	0.9	probability of resizing in DI
s_{pad}	0.9	scale of resizing in DI
N_{SI}	3	number of input decay in SI

TABLE 4. The default parameter settings of Adv-Makeup.

Name	Default	Description
α_1	1	weight of L_{attack}
α_2	1	weight of L_{gen}
β_1	0.1	weight of L_{grad}
β_2	0.1	weight of $L_{content}$
β_3	0.1	weight of L_{style}
lr	10^{-3}	learning rate of SGD optimizer
μ	0.9	momentum weight of SGD optimizer

302 high quality after-makeup faces. We select 195 after-makeup images with eyeshadow as makeup templates for makeup generation. 100 source before-makeup faces and 10 target before-makeup faces are selected randomly to form 1000 comparisons for impersonation attacks. (2) MT dataset [57], which consists of 1115 before-makeup facial images and 2719 makeup facial images with general quality. We randomly choose a total of 1000 comparisons similarly to evaluate the attack performance. For both datasets, the ASR results in this paper report the average results across 1000 identity pairs.

4) DEEP FACE RECOGNITION MODELS

The pre-trained FR models are divided into two parts: white-box surrogate models and black-box models. In our experiments, IR50, Resnet50, CosFace, and Mobilenet are used as surrogate models, and the pre-trained parameters are obtained from Face Robustness Benchmark (RobFR) [56]. Fowllowing [21], IR152 [58], IRSE50 [59], Facenet [6], MobileFace [1] are selected as black-box victim models to evaluate the attack transferability of above methods.

5) EVALUATION METRICS

We use the attack success rate (ASR) of black-box victim models to evaluate attack transferability. The ASR of model M is calculated as follows:

$$\mathrm{ASR_M} = \frac{\sum_{i=1}^{N_c} 1_\tau[\cos(M(I_{adv}^i), M(I_t^i)) > \tau]}{N_c} \times 100. \quad (18)$$

where 1_τ denotes the indicator function. The value of τ is set as the threshold at 0.01 FAR (False Acceptance Rate) for each victim FR model as most face recognition works do. τ of each victim model is: IR152 (0.167), IRSE50 (0.241), MobileFace (0.302), and Facenet (0.409). Meanwhile, we leverage FID [60], GMSD [61], and DISTS [62] to evaluate generated images' quality. FID, which is often used to evaluate the quality of images generated by GANs, measures the difference between the generated images and natural images from the perspective of image distribution. GMSD is based on the image gradient field to measure the similarity of two images. DISTS, which achieves outstanding results in image restoration and image super-resolution the as loss function [63], extracts image features by a pre-trained CNN and compares the structure of corresponding feature maps.

B. COMPARISION STUDY

1) EVALUATIONS ON ATTACK TRANSFERABILITY

In Table 5, every column represents the impersonation attack success results against the target black-box model under the corresponding dataset. The dark colored part indicates the best result and the light colored part indicates the second best result. Our proposed Adv-Eye outperforms the competitors under most attack conditions (slightly lower than Adv-Makeup against IR152 model under LADN dataset, slightly lower than Adv-Makeup against IRSE50, and lower than DI-CI-SI-MI-FGSM against Facenet under MT dataset). From the results, compared to other benchmark methods, Adv-Eye has the best attack transferability.

2) EVALUATION ON IMAGE QUALITY

As shown in Table 6, Adv-Eye outperforms Adv-Makeup in all quantitative evaluations. Comparing Figure 6(d) and Figure 6(e), Adv-Eye better preserves the semantic features of the source image (*e.g.* eyebrow shape and eyelashes) and generates eyeshadow with better visual quality. Both the numerical results and image results indicate that our method better arranges the perturbations and possesses stronger imperceptibility. In addition, BIM seems to achieves the best results in most of the numerical metrics, however, this is because BIM adds adversarial noise to the image and simply constrains the adversarial samples close to the original image based on pixel values. It can be shown in Figure 6. (b) that such noise perturbation does not fit the style of the source image, the visual quality of the image is low. What is more, BIM has poor attack transferability.

C. ABLATION STUDIES

1) IMAGE QUALITY

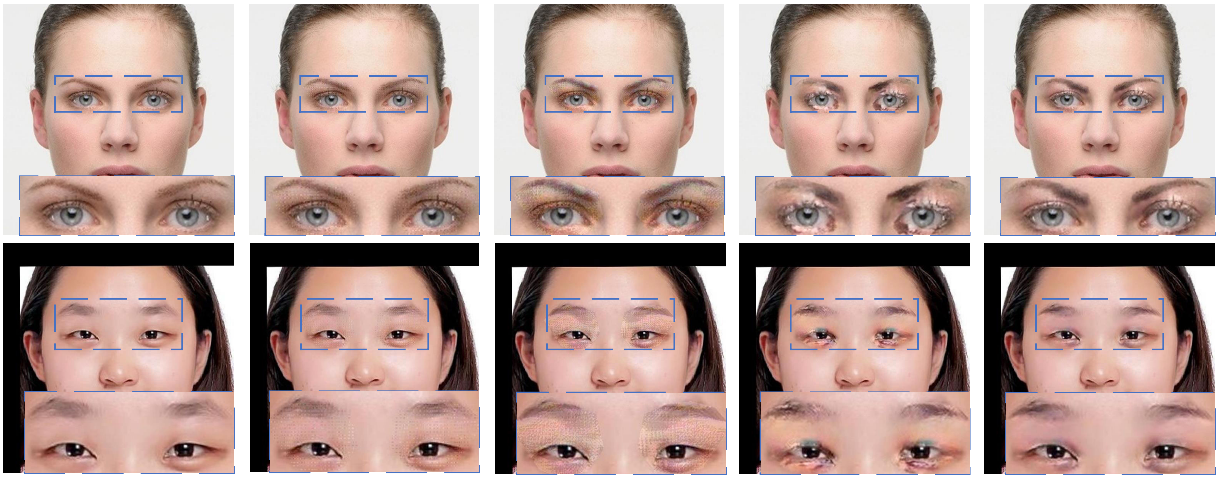
To illustrate the effectiveness of pre-makeup method and Makeup Blending Module used on improving image quality of the generated adversarial sample. We conduct an ablation experiment against them under LADN dataset and compare their numerical results and visualization results. As shown in Table 7, both methods are able to improve the image quality of the generated makeup adversarial samples. Moreover, the

TABLE 5. ASR (%) of impersonation attack over LADN dataset and MT dataset against IR152, IRSE50, Facenet and MobileFace.

Dataset	LADN Dataset				MT Dataset			
Target Model	IR152	IRSE50	FaceNet	MobileFace	IR152	IRSE50	FaceNet	MobileFace
BIM	17.50	41.50	8.20	58.80	29.70	72.20	49.00	82.70
DI-CI-SI-MIM	21.20	43.50	24.90	61.20	33.40	71.80	78.00	82.30
Adv-Makeup	30.78	59.11	31.00	70.56	33.50	73.20	73.00	81.40
Adv-Eye	29.78	59.33	34.00	74.89	34.50	73.00	74.70	83.10

TABLE 6. Results of image quality evaluation under LADN dataset and MT dataset. Each column indicates the result of each comparison method under the corresponding metric.

Dataset	LADN Dataset			MT Dataset		
Numerical indicators	FID [60]	GMSD [61]	DISTS [62]	FID [60]	GMSD [61]	DISTS [62]
BIM	4.594	0.0163	0.0149	6.534	0.0186	0.0207
DI-CI-SI-MIM	10.237	0.0385	0.0181	18.994	0.0456	0.0279
Adv-Makeup	13.948	0.0687	0.0230	12.678	0.0669	0.0251
Adv-Eye	6.425	0.0478	0.0173	6.186	0.0444	0.0176



(a) Soure Image (b) BIM (c) DI-CI-SI-MIM (d) Adv-Makeup (e) Adv-Eye

FIGURE 6. Adversarial samples and eyes' amplifications generated by comparision methods in LADN dataset (first row) and MT dataset (second row).

TABLE 7. Image quality results of ablation experiments of using pre-makeup images and the makeup blending module.

	FID [60]	GMSD [61]	DISTS [62]
W.O. pre-makeup	9.548	0.0566	0.0188
W.O. Makeup Blending	8.825	0.0548	0.0186
Adv-Eye	**6.425**	**0.0478**	**0.0173**

TABLE 8. ASR results of using the attack method of Adv-makeup and using attack module.

	IR125	IRSE50	FaceNet	MobileFace
Adv-Makeup	**30.89**	56.88	31.56	72.00
Ours	29.78	**59.33**	**34.00**	**74.89**

combination of them can further enhance the image quality. Furthermore, as shown in Figure 7, by using pre-makeup image as the ground truth of discriminator D, generator G can be more effectively guided to generate eye makeup and reduce the modification of other semantic content. By introducing Makeup Blending Module, the generated eye makeup can be more natural and closer in style to the source image.

2) ATTACK TRANSFERABILITY

As mentioned above, relative to the attack method proposed by Adv-Makeup, Attack Module more fully uses white-box surrogate models' information, thereby enhancing the black-box transferability. To further illustrate this point, we compare the ASR of four black-box FR models of the Adv-Eye model, which are trained using the attack method proposed by Adv-Makeup and Attack Module under LADN dataset.

As shown in Table 8, except for IR152, using Attack Module yields higher ASRs. The results indicate that Attack Module is able to enhance the attack transferability of the generated adversarial samples.

D. ATTACK PERFORMANCE ON COMMERCIAL APIS

To test the attack effectiveness of Adv-Eye in practical application scenarios, we introduce two widely used commercial

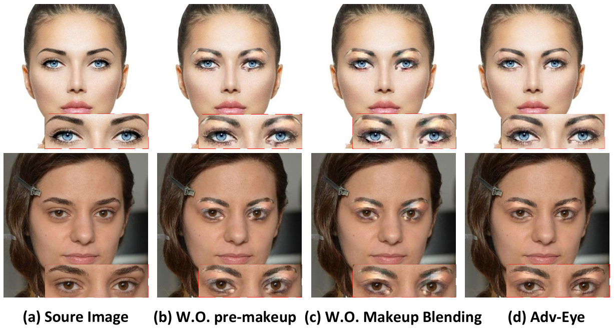


FIGURE 7. Illustration of adversarial samples and eyes' amplifications generated under each ablation setting.

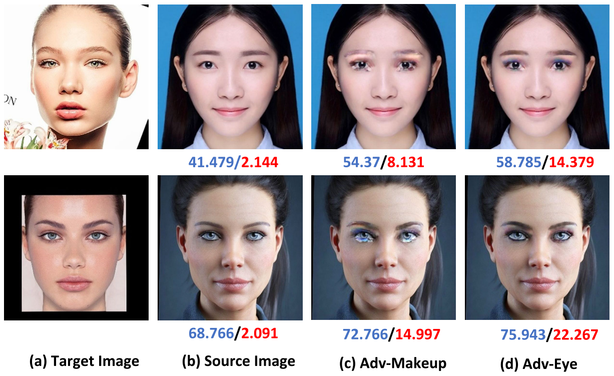


FIGURE 8. Illustration of recognition results of online commercial face recognition models. the numbers below the image are confidence results of commercial FR platforms (red: Aliyun, blue: Face++).

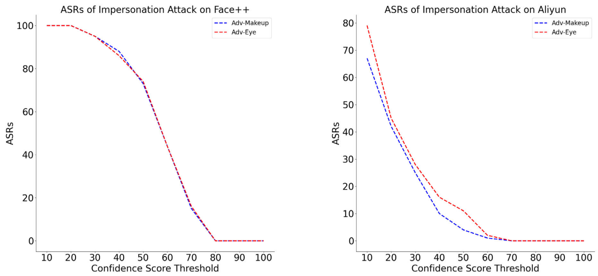


FIGURE 9. ASR comparison results of impersonation attacks of Adv-Eye and Adv-Makeup along with the confidence score threshold changes on Face++ and Aliyun. (red: Adv-Eye, blue: Adv-Makeup).

face recognition platforms Aliyun and Face++. The attack effectiveness under two platforms is expressed as the change in confidence scores after adding adversarial eye makeup. As shown in Figure 8, compared to Adv-Makeup, the makeup adversarial samples generated by Adv-Eye achieve higher confidence in both platforms while maintaining higher attack imperceptibility. Additionally, taking the same target as shown in the first row of Figure 8, Figure 9 shows the attack success rate of Adv-Makeup and Adv-Eye against Face++ and Aliyun at different confidence thresholds. Figure 9 shows that Adv-Eye achieves better ASR in most threshold conditions on face++ and Aliyun.

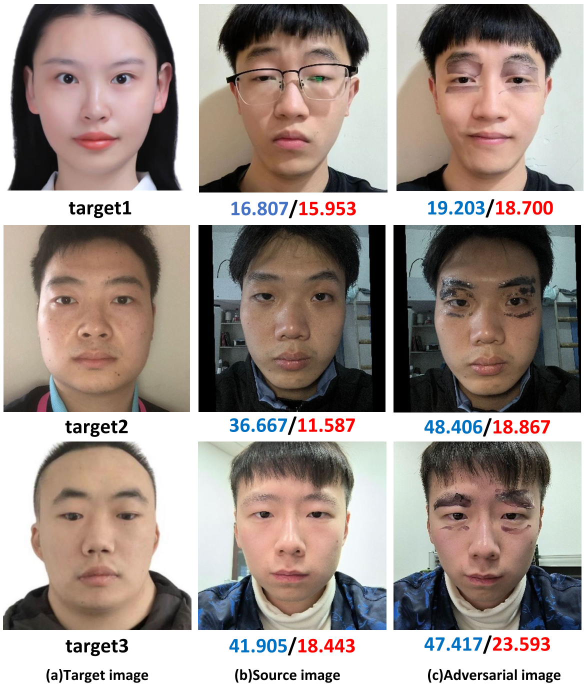


FIGURE 10. Illustration of attack results against commercial platforms in realistic scenarios by pasting tatoo paster. The numbers below the image are confidence results of commercial FR platforms (red: Aliyun, blue: Face++).

E. ATTACK PERFORMANCE IN REAL-WORLD

To further investigate the usefulness of Adv-Eye in realistic scenarios, we physically implement adversarial eyeshadow using simple tatoo paster. Volunteers attack Face++ and Aliyun by pasting the corresponding eyeshadow stickers over orbital region. As shown in Figure 10, although the error caused during sticker printing and cutting have a certain impact on attack effect, the eyeshadow generated by Adv-Eye is still able to improve the confidence scores of both systems. This shows that Adv-Eye has the facial recognition systems in the real world.

V. CONCLUSION

In this paper, we present Adv-Eye, a novel model that aims to deceive deep face recognition (FR) systems by adding indiscernible eyeshadow to facial images. In Makeup Generation Module, we propose the pre-makeup method to help GANs accurately edit the eye area. We propose Makeup Blending Module to evaluate adversarial samples for naturalness from multiple angles, reducing the visual artifacts caused by fusing eyeshadow to the source image. Moreover, we propose Attack Module which efficiently utilizes white-box substitute FR models by using an ensemble attack strategy based on meta-learning and input decay. Compared to the attack methods which have same type of perturbations, Adv-Eye achieves the best FID and DISTS scores, 6.186 and 0.0176 respectively and outperforms in terms of visualization effects. Besides, Adv-Eye demonstrates superior attack capabilities, as it shows an average increase of 1.63% in attack success rate on local FR models and an average boost of 5.33 in the confidence score on online face recognition platforms. Adv-Eye achieves a better balance between the visual quality and attack effects of adversarial samples. Thus, Adv-Eye represents a more significant threat to face recognition models under practical conditions. Nonetheless, Adv-Eye has limitations, Adv-Eye has limited improvement in attack capability and the generated adversarial eyeshadow has insufficient robustness. As shown in Figure 10, due to inevitable errors when applying the eyeshadow in real-world setting, the naturalness and attack transferability of adversarial eyeshadow will be significantly reduced. In the future, we will further study the differences between different FR models and design black-box attack strategies that can more elaborately use white-box model information. In addition, we will consider the error that may occur when implementing perturbations under real-world conditions, in order to better maintain the attack effect of generated adversarial samples.

ACKNOWLEDGMENT

(Jiatian Pi and Junyi Zeng contributed equally to this work.)

REFERENCES

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[2] M. I. Sharif, M. A. Khan, A. Alqahtani, M. Nazir, S. Alsubai, A. Binbusayyis, and R. Damaševičius, "Deep learning and kurtosis-controlled, entropy-based framework for human gait recognition using video sequences," *Electronics*, vol. 11, no. 3, p. 334, Jan. 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/3/334

[3] M. J. Umer and M. I. Sharif, "A comprehensive survey on quantum machine learning and possible applications," *Int. J. E-Health Med. Commun.*, vol. 13, no. 5, pp. 1–17, Dec. 2022, doi: 10.4018/IJEHMC.315730.

[4] C. Luo, L. Yang, G. Zhao, N. Jiang, J. Pi, and Z. Wu, "Mixture of experts for facial forgery detection," *J. Imag. Sci. Technol.*, vol. 66, no. 6, 2022, Art. no. 060501.

[5] R. Zahra, A. Shehzadi, M. I. Sharif, A. Karim, S. Azam, F. D. Boer, M. Jonkman, and M. Mehmood, "Camera-based interactive wall display using hand gesture recognition," *Intell. Syst. With Appl.*, vol. 19, Sep. 2023, Art. no. 200262. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266730532300087X

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

[7] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "CurricularFace: Adaptive curriculum learning loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5900–5909.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[10] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[11] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7706–7714.

[12] L. Yang, Q. Song, and Y. Wu, "Attacks on state-of-the-art face recognition using attentional adversarial attack generative network," *Multimedia Tools Appl.*, vol. 80, no. 1, pp. 855–875, Jan. 2021.

[13] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2017, *arXiv:1712.09665*.

[14] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," *ACM Trans. Privacy Secur.*, vol. 22, no. 3, pp. 1–30, Aug. 2019.

[15] S. Komkov and A. Petiushko, "AdvHat: Real-world adversarial attack on ArcFace face ID system," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 819–826.

[16] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces: Adversarial face synthesis," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–10.

[17] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, *arXiv:1607.02533*.

[18] S. Jia, B. Yin, T. Yao, S. Ding, C. Shen, X. Yang, and C. Ma, "Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NeruIPS)*, vol. 35, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. Red Hook, NY, USA: Curran Associates, 2022, pp. 34136–34147. https://proceedings.neurips.cc/paper_files/paper/2022/file/dccbeb7a8df3065c4646928985edf435-Paper-Conference.pdf

[19] C.-S. Lin, C.-Y. Hsu, P.-Y. Chen, and C.-M. Yu, "Real-world adversarial examples via makeup," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2854–2858.

[20] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. Wu, "Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14994–15003.

[21] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1252–1258.

[22] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, and C. Xie, "Adversarial attacks and defences competition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*. Cham, Switzerland: Springer, 2018, pp. 195–231.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[24] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, Apr. 2017, pp. 506–519.

[25] F. Waseda, S. Nishikawa, T.-N. Le, H. H. Nguyen, and I. Echizen, "Closer look at the transferability of adversarial examples: How they fool different models differently," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1360–1368.

[26] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2725–2734.

[27] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4307–4316.

[28] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=SJlHwkBYDH

[29] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the transferability of adversarial attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16138–16147.

[30] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[31] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1924–1933.

[32] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017. [Online]. Available: https://openreview.net/forum?id=Sys6GJqxl

[33] X. Kang, B. Song, X. Du, and M. Guizani, "Adversarial attacks for image segmentation on multiple lightweight models," *IEEE Access*, vol. 8, pp. 31359–31370, 2020.

[34] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, ''Adversarial t-shirt! evading person detectors in a physical world,'' in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 665–681.

[35] V. Cherepanova, M. Goldblum, H. Foley, S. Duan, J. P. Dickerson, G. Taylor, and T. Goldstein, ''LowKey: Leveraging adversarial attacks to protect social media users from facial recognition,'' in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=hJmtwocEqzc

[36] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, ''The unreasonable effectiveness of deep features as a perceptual metric,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[37] Z. Xiao, X. Gao, C. Fu, Y. Dong, W. Gao, X. Zhang, J. Zhou, and J. Zhu, ''Improving transferability of adversarial patches on face recognition with generative models,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11840–11849.

[38] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, ''Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,'' in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 1528–1540.

[39] T. Karras, S. Laine, and T. Aila, ''A style-based generator architecture for generative adversarial networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[40] Z.-A. Zhu, Y.-Z. Lu, and C.-K. Chiang, ''Generating adversarial examples by makeup attacks on face recognition,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2516–2520.

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ''Generative adversarial nets,'' in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 27, 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[42] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, ''Image-to-image translation with conditional adversarial networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, ''Unpaired image-to-image translation using cycle-consistent adversarial networks,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[44] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, ''Meta-learning in neural networks: A survey,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.

[45] R. Ni, M. Goldblum, A. Sharaf, K. Kong, and T. Goldstein, ''Data augmentation for meta-learning,'' in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, M. Meila and T. Zhang, Eds., Jul. 2021, pp. 8152–8161. [Online]. Available: https://proceedings.mlr.press/v139/ni21a.html

[46] J. Shin, H. B. Lee, B. Gong, and S. J. Hwang, ''Large-scale meta-learning with continual trajectory shifting,'' in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, M. Meila and T. Zhang, Eds., Jul. 2021, pp. 9603–9613. [Online]. Available: https://proceedings.mlr.press/v139/shin21a.html

[47] Z. Yuan, J. Zhang, Y. Jia, C. Tan, T. Xue, and S. Shan, ''Meta gradient adversarial attack,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7728–7737.

[48] H. Chang, J. Lu, F. Yu, and A. Finkelstein, ''PairedCycleGAN: Asymmetric style transfer for applying and removing makeup,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 40–48.

[49] P. Pérez, M. Gangnet, and A. Blake, ''Poisson image editing,'' *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, Jul. 2003, doi: 10.1145/882262.882269.

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, ''Image quality assessment: From error visibility to structural similarity,'' *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[51] L. Gatys, A. S. Ecker, and M. Bethge, ''Texture synthesis using convolutional neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 28, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/a5e00132373a7031000fd987a3c9f87b-Paper.pdf

[52] L. Zhang, T. Wen, and J. Shi, ''Deep image blending,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 231–240.

[53] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang, ''LADN: Local adversarial disentangling network for facial makeup and de-makeup,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10480–10489.

[54] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015.

[55] A. Kurakin, I. Goodfellow, and S. Bengio, ''Adversarial machine learning at scale,'' 2016, *arXiv:1611.01236*.

[56] X. Yang, D. Yang, Y. Dong, H. Su, W. Yu, and J. Zhu, ''RobFR: Benchmarking adversarial robustness on face recognition,'' 2020, *arXiv:2007.04118*.

[57] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin, ''BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network,'' in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 645–653, doi: 10.1145/3240508.3240618.

[58] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[59] J. Hu, L. Shen, and G. Sun, ''Squeeze-and-excitation networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[60] 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. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

[61] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, ''Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,'' *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[62] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, ''Image quality assessment: Unifying structure and texture similarity,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.

[63] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, ''Comparison of full-reference image quality models for optimization of image processing systems,'' *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1258–1281, Apr. 2021.

JIATIAN PI received the B.S. degree in communication engineering from Chongqing University, Chongqing, China, in 2012, and the Ph.D. degree in information and communication engineering from the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China, in 2017. Since 2017, he has been with the National Center for Applied Mathematics in Chongqing, Chongqing Normal University. He has authored more than 15 articles and seven inventions. His research interests include computer vision, machine learning, and adaptive traffic signal control.

JUNYI ZENG received the bachelor's degree in mathematics from the Southwest University of Science and Technology, in 2020. He is currently pursuing the master's degree in computational mathematics with Chongqing Normal University. His research interests include optimization algorithm, computer vision, and adversarial attack.

QUAN LU received the Ph.D. degree in operations research from the University of Southern California. He is currently the Head of the Mashang Consumer Finance Artificial Intelligence Research Institute. Before joining Mashang, he was the Senior Director of the Data Science Team, Alibaba and Yahoo! respectively.

NING JIANG received the bachelor's degree from Kyoto University, and the master's degree from The University of Tokyo. He is currently the Deputy General Manager and the Chief Information Officer of Mashang Consumer Finance Company Ltd. He has more than 20 years of practical experience in the field of internet architecture and financial technology, and has led the development of a 10 million QPS internet online transaction system and an industry-leading core financial transaction system. Under his excellent research and development management and innovation leadership, Mashang has built more than 900 systems covering the entire business process of retail finance, applied for more than 600 invention patents, and it is the first financial institution in Chongqing to be certified as a ''National High-Tech Enterprise'' and has been ranked among the ''KPMG Top 50 Leading Financial Technology Enterprises in China'' for six consecutive years.

HAIYING WU received the bachelor's degree in computer science and the master's degree in computer software from the University of Science and Technology of China, in 1995 and 1998, respectively. From 1998 to 2003, he was a Research and Development Engineer with Shanghai Wanda Information Company Ltd., and the CTO of the Just System Shanghai Research and Development Center. From 2003 to 2019, he was the Senior Technical Director with Hewlett-Packard Company Ltd., China, and the product line Vice President with Huawei Technologies Company Ltd. He is currently the Senior Technical Director of the Financial Technology Research and Development Department, Mashang Consumer Finance Company Ltd.

LINCHENGXI ZENG received the bachelor's degree in international business management and the master's degree in software engineering from the Dalian University of Technology, China, in 2002 and 2004, respectively. From 2005 to 2016, he was an Architect with China Dalian Hewlett-Packard Company Ltd., and PINGAN Technology Company Ltd. He was the Chief Technology Officer with Shenzhen Tianhong Information Technology Company Ltd. He is currently the Head of the System Architecture Department, Mashang Consumer Finance Company Ltd., Beijing.

ZHIYOU WU received the B.Sc. degree from Chongqing Normal University, in 1987, and the M.Sc. and Ph.D. degrees from the Department of Mathematics, Shanghai University, in 1990 and 2003, respectively. She was with Chongqing Normal University. She has published more than 80 papers in various international journals, including *SIAM Journal on Optimization*, *Nonlinear Analysis*, *Journal of Global Optimization*, and *Journal of Optimization Theory and Applications*. She has successfully secured more than 20 research projects, including grants from the US ARC, Humboldt Fellowship, Chinese National Natural Science Foundation, and other important projects in China. Her research interests include optimization, global optimization theory and algorithms, integer programming, and nonlinear programming.