Adversarial Relighting Against Face Recognition

Qian Zhang, Qing Guo, *Member, IEEE*, Ruijun Gao, Felix Juefei-Xu, *Member, IEEE*, Hongkai Yu, *Member, IEEE*, and Wei Feng, *Member, IEEE*

***Abstract*— Deep face recognition (FR) has achieved significantly high accuracy on several challenging datasets and fosters successful real-world applications, even showing high robustness to the illumination variation that is usually regarded as a main threat to the FR system. However, in the real world, illumination variation caused by diverse lighting conditions cannot be fully covered by the limited face dataset. In this paper, we study the threat of lighting against FR from a new angle, *i.e.*, *adversarial attack*, and identify a new task, *i.e.*, *adversarial relighting*. Given a face image, adversarial relighting aims to produce a naturally relighted counterpart while fooling the state-of-the-art deep FR methods. To this end, we first propose the physical model-based adversarial relighting attack (ARA) denoted as *albedo-quotient-based adversarial relighting attack (AQ-ARA)*. It generates natural adversarial lighting under the guidance of FR systems and synthesizes adversarially relighted face images. Moreover, we propose the *auto-predictive adversarial relighting attack (AP-ARA)* by training an adversarial relighting network (ARNet) to automatically predict the adversarial lighting in a one-step manner according to different input faces, allowing efficiency-sensitive applications. More importantly, we propose to transfer the above digital attacks to *physical ARA (Phy-ARA)* through a precise relighting device, making the estimated adversarial lighting condition reproducible in the real world. We validate our methods on several state-of-the-art deep FR methods on two public datasets. The extensive and insightful results demonstrate our work can generate realistic adversarial relighted face images fooling face recognition tasks easily, revealing the threat of specific light directions and strengths.**

***Index Terms*— Adversarial relighting, adversarial attack, face recognition.**

Manuscript received 28 January 2023; revised 28 September 2023 and 12 December 2023; accepted 29 January 2024. Date of publication 22 March 2024; date of current version 3 October 2024. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62072334. The work of Qing Guo was supported in part by the National Research Foundation, Singapore; in part by Defence Science Organisation (DSO) National Laboratories under the AI Singapore Program (AISG) under Award AISG2-GC-2023-008; in part by the Career Development Fund (CDF) of Agency for Science, Technology and Research (A*STAR) under Grant C233312028; and in part by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative under Grant DTC-RGC-04. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christian Riess. *(Qian Zhang and Qing Guo are co-first authors.) (Corresponding author: Wei Feng.)*

Qian Zhang, Ruijun Gao, and Wei Feng are with the School of Computer Science and Technology, College of Intelligence and Computing, Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin 300305, China, and also with the Key Research Center for Surface Monitoring and Analysis of Cultural Relics (SMARC), State Administration of Cultural Heritage, Beijing 100009, China (e-mail: wfeng@ieee.org).

Qing Guo is with the Institute of High Performance Computing (IHPC) and the Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore 138632.

Felix Juefei-Xu is with the Tandon School of Engineering, New York University, New York, NY 10003 USA.

Hongkai Yu is with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, OH 44115 USA.

Digital Object Identifier 10.1109/TIFS.2024.3380848

I. Introduction

The fast-paced development of deep learning (DL) has bolstered the deployment of high-performance DL-based face recognition systems (FRS) [1], [2]. Compared to non-DL-based FRS from a decade ago, the DL-based FRS nowadays can handle more challenging unconstrained scenarios and is very well suited for handling various FR tasks in unconstrained real-world scenarios, especially when faces are under various known or unknown degradations such as being very low-resolution [3], [4], [5], at an off-angle pose [6], [7], [8], heavily occluded by objects or crowd [9], [10], [11], *etc*. Among the mentioned degradation factors, illumination variation is one of the most challenging due to its high variability and pervasiveness for faces due to the collective effects of complex environmental lighting as well as the facial structure and reflectance properties. Illumination variation robustness of face recognition systems is very important for many real-world applications (e.g., surveillance [12], authentication [13], marketing [14]), especially for the outdoor environment, since the lighting conditions cannot be controlled.

To mitigate the challenges that illumination variations have posed on FRS, face relighting was proposed to adjust the lighting appearance on a given face image. More specifically, face relighting aims at altering how light and shadow are cast on the face based on a desired illumination that is usually a result of diffuse or directional lighting. Traditionally, face relighting requires the decomposition of the face image into shape geometry, lighting, and reflectance maps, respectively, and then novel lighting, *i.e.*, relighting, is achieved by swapping the intrinsic lighting map with the desired one. However, the accuracy of such relighting and its realisticity depends on the precise estimation of the geometry and reflectance map, which is a difficult task to accomplish, especially when the faces are not in an ideal studio setting, but instead in an unconstrained environment. DL-based face relighting methods attempt to execute the aforementioned process in an end-to-end or semi-end-to-end fashion by leveraging large-scale training image pairs. In these recent studies [15], [16], [17], [18], we have seen a jump in face relighting performance both quantitatively and qualitatively, compared to traditional non-DL-based relighting methods.

By capitalizing the advances in DL-based face relighting capabilities, in this work, we are proposing a new study that aims at revealing the vulnerabilities of FRS from the angle of face relighting. To be more specific, we have identified a new task, *i.e.*, adversarial face relighting attack, whose goal is, given a source face image, to produce a naturally relighted counterpart while fooling the state-of-the-art deep FR

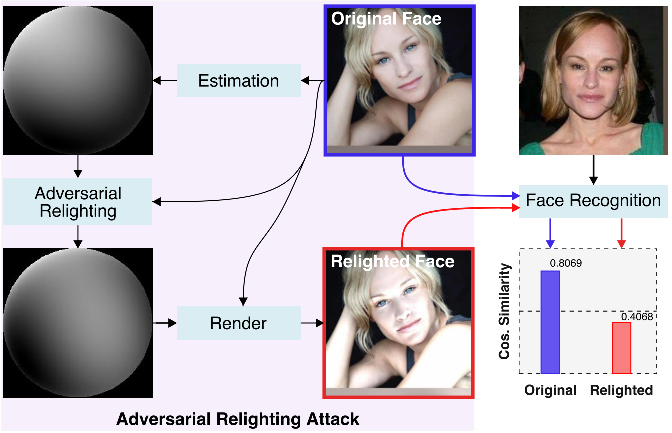


Fig. 1. Intuitive idea of the new task *adversarial relighting attack (ARA)*. The original face is relighted via the ARA, making the face recognition system fail to identify the same person, that is, the cosine similarity reduces from 0.8069 to 0.4068.

methods (See Fig. 1 for the intuitive idea). **First**, we propose the physical model-based adversarial relighting attack (ARA), that is, *albedo-quotient-based adversarial relighting attack (AQ-ARA)*. Specifically, we define an adversarial objective function based on the physical lighting model and tune the lighting parameters by maximizing the function. As a result, we can generate natural adversarial light under the said physical lighting model and the guidance of FRS and synthesize the adversarially relighted face images. **Second**, we design the auto-predictive adversarial relighting attack (AP-ARA) by training an adversarial relighting network (ARNet) to automatically predict the adversarial light in a one-step manner according to different input faces, allowing efficiency-sensitive applications. **Third**, we propose to transfer the aforementioned digital adversarial attacks to *physical ARA (Phy-ARA)* through a precise relighting device, making the estimated adversarial lighting condition reproducible in the real world. We validate our methods on three state-of-the-art deep FR methods on two public face recognition datasets. **More importantly**, we conduct both digital and physical experiments to analyze the effects of light to FR via our ARAs, revealing and validating the threat of challenging lighting conditions. To the best of our knowledge, this work is the very first attempt to study how face relighting can be capitalized to adversarially affect the FRS from the angle of the proposed adversarial relighting attacks.

II. Related Work

A. Relighting Methods

In the areas of computer vision and graphics, relighting is an effective way to adjust the illumination variations for an enhanced or different-style visualization. For example, [19] uses a spherical harmonic representation for the target lighting and train the network with a synthetic dataset created by relighting single portrait images using a traditional ratio image-based method. Their method can generate a relit portrait image by using a source image and a target lighting as input. Reference [20] proposes a differentiable specular rendering layer to render low-frequency non-Lambertian materials under various illuminations of spherical harmonic. Reference [21] proposes a self-supervised method for image relighting of single view images based on an auto-encoder network. Reference Besides, quotient-based relighting methods have also been studied widely. Reference [22] proposes to use the quotient image for portrait relighting. They re-render the input front-view image to simulate new illumination conditions given a sample of reference images with varying illumination conditions. Reference [17] extends the ratio image to arbitrary pose by aligning facial landmarks of the source and target image, then they can synthesize the re-illuminated face images in different pose angles. Reference [23] introduces a new method that uses the shadow mask to estimate the ambient light intensity in an image, their method learn to predict the quotient image between a source image and the target image with the desired lighting. Reference [16] proposes a novel framework that explicitly models multiple reflectance channels for single image portrait relighting, and this method can generate the photorealistic portrait relighting image based on a reference image or an environment map. Reference [15] produces a portrait relighting image with an encoder-decoder based CNN network where the target illumination is incorporated in the bottleneck of the network. Reference [18] provides a reinforcement learning-based approach to portrait relighting. Different with the existing methods, this paper is to design a new deep adversarial relighting method to attack the face recognition.

B. Attacks Against Face Recognition

The attacks against face recognition might make the current face recognition systems vulnerable. They could be roughly divided into four classes [24]: Perturbing, Morphing, Retouching, Tampering. Perturbing attack adds imperceptible perturbations so as to fool the face recognition system. Reference [25] shows that several deep learning based commercial face recognition algorithms and systems are vulnerable to different adversarial perturbation attacks. Morphing attack is to generate a morphed face with the imperceptible adversarial attack embedded for face recognition. References [26] and [27] shows that some of current commercial face recognition systems cannot protect users from morphed faces. The facial retouching attack is also possible to fool some commercial face recognition systems [28]. Recently, generator adversarial networks (GANs) based fake image synthesis and the well-known DeepFake method [29], [30] could generate or modify the face images, leading to the tampering based attack to face recognition. Different with the above methods, the proposed adversarial relighting method attacks the face recognition from a new perspective of face relighting.

C. General Adversarial Attacks

Recently, the adversarial attack to fool deep neural networks has attracted many research attentions. Given an image as input, the adversarial attack can be realized by adding imperceptible noises or applying natural transformations. On the one hand, several adversarial noise attack methods got promising attack results, such as gradient computation based fast gradient sign method (FGSM) [31], iterative-version FGSM [32], momentum iterative FGSM [33], different distance metrics based C&W method [34], attended regions and features based TAA (transferable attentive attack) method [35], randomization based [36], perceptually aware and stealthy adversarial denoise [37], and so on. On the other hand, some natural transformations that are imperceptible to humans can be applied for image attack. For example, the adversarial attacks can be implemented with various transformations, *e.g.*, semantic-aware colorization or texture transfer [38], motion blurring synthesis [39], watermark overlap [40], rain [41] and haze synthesis [42], *etc*. By adding the adversarial relighting for attack, the adversarial relighting method proposed in this paper is a kind of novel natural transformation based adversarial attack method.

III. Adversarial Relighting Attack

In this section, we propose two adversarial relighting attacking (ARA) methods. The first one is based on the physical model and denoted as the albedo-quotient-based ARA (AQ-ARA) (see Section III-A). The second one uses the CNN to automatically predict the adversarial light in a one-step way (*i.e.*, auto-predictive-based ARA (AP-ARA)) (see Section III-B). With these two methods, we further design a physical ARA Section III-C to reproduce the adversarial light in the real world.

A. Albedo-Quotient-Based ARA (AQ-ARA)

Given a face image $\mathbf{I}$, we assume it follows the Lambertian model that is a widely used face rendering model. Thus, we can represent the face image as

$$\mathbf{I} = \mathbf{R} \odot \mathrm{f}(\mathbf{N}, \mathbf{L}), \tag{1}$$

where $\mathbf{R}$, $\mathbf{N}$, and $\mathbf{L}$ denote the reflectance, normal, and lighting (spherical harmonic coefficients form), respectively. $\mathrm{f}(\cdot)$ is the spherical harmonic function. More specifically, $\mathbf{L}$ is a nine-dimensional vector corresponding to the nine spherical harmonic coefficients. Let $\mathrm{g}(\mathbf{N}_i, \mathbf{L})$ be the spherical harmonic function for one pixel, i.e., $\mathrm{f}(\mathbf{N}, \mathbf{L}) = [\mathrm{g}(\mathbf{N}_1, \mathbf{L}), \mathrm{g}(\mathbf{N}_2, \mathbf{L}), \cdots, \mathrm{g}(\mathbf{N}_i, \mathbf{L}), \cdots]^{\mathrm{T}}$, where $\mathbf{N}_i$ denotes the normal of i-th pixel. Then function $\mathrm{g}(\mathbf{N}_i, \mathbf{L})$ has the following form,

$$\mathrm{g}(\mathbf{N}_i, \mathbf{L}) = \sum_{k=0}^{8} \mathbf{L}_k \mathrm{H}_k(\mathbf{N}_i), \tag{2}$$

where $\mathbf{L} = [\mathbf{L}_0, \cdots, \mathbf{L}_8]^{\mathrm{T}}$. Let $\mathbf{N}_i = [\mathbf{N}_i^{\mathrm{x}}, \mathbf{N}_i^{\mathrm{y}}, \mathbf{N}_i^{\mathrm{z}}]^{\mathrm{T}}$, then we have

$$\begin{aligned} &\mathrm{H}_0 = 1.0, \quad \mathrm{H}_1(\mathbf{N}_i) = \mathbf{N}_i^{\mathrm{y}}, \quad \mathrm{H}_2(\mathbf{N}_i) = \mathbf{N}_i^{\mathrm{z}}, \\ &\mathrm{H}_3(\mathbf{N}_i) = \mathbf{N}_i^{x}, \; H_4(\mathbf{N}_i) = \mathbf{N}_i^{\mathrm{x}}\mathbf{N}_i^{\mathrm{y}}, \quad H_5(\mathbf{N}_i) = \mathbf{N}_i^{\mathrm{y}}\mathbf{N}_i^{\mathrm{z}}, \\ &H_6(\mathbf{N}_i) = -\mathbf{N}_i^{\mathrm{x}}\mathbf{N}_i^{\mathrm{x}} - \mathbf{N}_i^{\mathrm{y}}\mathbf{N}_i^{\mathrm{y}} + 2\mathbf{N}_i^{\mathrm{z}}\mathbf{N}_i^{\mathrm{z}}, \\ &H_7(\mathbf{N}_i) = \mathbf{N}_i^{\mathrm{z}}\mathbf{N}_i^{\mathrm{x}}, \quad H_8(\mathbf{N}_i) = \mathbf{N}_i^{\mathrm{x}}\mathbf{N}_i^{\mathrm{x}} - \mathbf{N}_i^{\mathrm{y}}\mathbf{N}_i^{\mathrm{y}}. \end{aligned} \tag{3}$$

Our objective is to update the lighting (*i.e.*, $\mathbf{L}$) to a new one (*i.e.*, $\hat{\mathbf{L}}$) and produce a new face $\hat{\mathbf{I}}$ that can mislead a state-of-the-art face recognition (FR) method. We denote $\hat{\mathbf{L}}$ as the *adversarial lighting*. This new task actually combines the face relighting with the adversarial attack, thus we name it as the *adversarial relighting attack*. According to Eq. (1), we need to estimate the reflectance and normal, which is not easy since calculating the accurate reflectance map is still an open problem. To alleviate the requirement, we borrow the albedo-quotient image introduced in [43] to get the reflectance-free method for relighting. Specifically, we can represent the same face with different lighting conditions as $\mathbf{I} = \mathbf{R} \odot \mathrm{f}(\mathbf{N}, \mathbf{L})$ and $\hat{\mathbf{I}} = \mathbf{R} \odot \mathrm{f}(\mathbf{N}, \hat{\mathbf{L}})$. As a result, we have

$$\hat{\mathbf{I}} = \mathbf{R} \odot \mathrm{f}(\mathbf{N}, \hat{\mathbf{L}}) = \frac{\mathrm{f}(\mathbf{N}, \hat{\mathbf{L}})}{\mathrm{f}(\mathbf{N}, \mathbf{L})}\mathbf{I}. \tag{4}$$

With Eq. (4), we can relight the face image $\mathbf{I}$ through the normal $\mathbf{N}$, the original lighting $\mathbf{L}$, and the targeted light $\hat{\mathbf{L}}$ [19]. We first estimate face normal $\mathbf{N}$ and lighting $\mathbf{L}$ following the implementation of [19] and [44]. Then we define the objective function to estimate the adversarial lighting. Given a deep neural network $\phi(\cdot)$ for the FR, we aim to calculate the adversarial lighting $\hat{\mathbf{L}}$ to let the relighted face image $\hat{\mathbf{I}}$ mislead the FR by solving

$$\begin{aligned} \hat{\mathbf{L}} = &\arg\min_{\mathbf{L}'} \cos\Big(\phi\Big(\frac{\mathrm{f}(\mathbf{N}, \mathbf{L}')}{\mathrm{f}(\mathbf{N}, \mathbf{L})}\mathbf{I}\Big), \phi(\mathbf{I})\Big), \\ &\text{subject to } \|\mathbf{L}' - \mathbf{L}\|_\infty \le \epsilon \end{aligned} \tag{5}$$

where $\cos(\cdot)$ denotes the cosine similarity function and $\phi(\cdot)$ is the embedding of the input face images. Parameter ϵ controls the changing degrees of lighting. Intuitively, we minimize the similarity by tuning the lighting, that is, to let the relighted face be different from the original counterpart under the constraint of ϵ. We can calculate the gradient of the loss function with respect to light to realize the gradient-based attack. As a result, the attack method can be integrated into any gradient-based additive-perturbation attack methods, *e.g.*, FGSM, BIM, MIFGSM. Here, we use the sign gradient descent optimization with the step size $\lambda = \frac{\epsilon}{T}$. T denotes the iteration number and we fix it as ten as a common setup in adversarial attacks. We show an example of AQ-ARA in Fig. 2. Compared with the random relighting, the adversarial relighting lets the similarity with the reference image decrease significantly while having realistic appearance. The random relighting means we uniformly sample a light $\mathbf{L}$ within the range of $[-\epsilon, \epsilon]$ and relight the face via Eq. (4).

B. Auto-Predictive-Based ARA (AP-ARA)

Although above method is able to achieve effective adversarial relighting, the iterative optimization manner limits the potential applications in particular for efficiency-sensitive applications. In the following, we propose an adversarial relighting network (ARNet) to adaptively predict the adversarial lighting in an one-step way.

Given the original lighting $\mathbf{L}$, we use a deep neural network to map the $\mathbf{L}$ to the adversarial lighting $\hat{\mathbf{L}}$, directly. To realize this goal, a naive solution is to build the lighting pair dataset through the method introduced in Section III-A. Nevertheless, this strategy might cost a lot of time due to the iteration operation. To alleviate this problem, we propose an end-to-end network denoted as adversarial relighting network (ARNet)

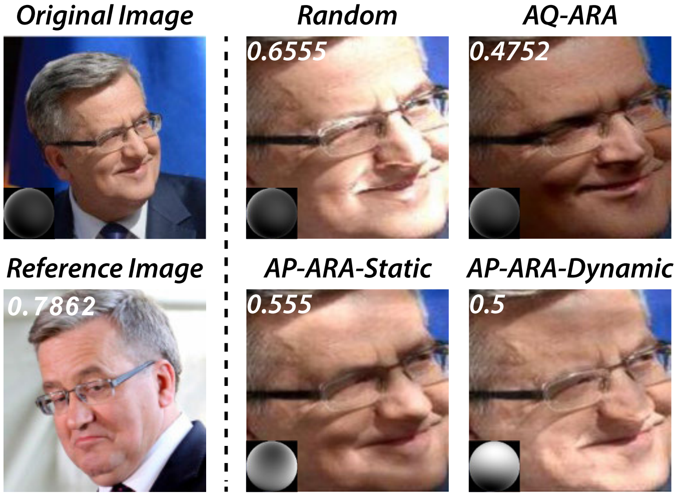


Fig. 2. An relighting example of Random, AQ-ARA, and AP-ARA (both static and dynamic). The cosine similarity between each relighted face and the reference face based on FaceNet are showed at the left-hand corner.

containing three stages, *e.g.*, predicting the original light, estimating the adversarial lighting based on the original one, and relighting the face under the adversarial lighting. We can train this network under the supervision of cosine similarity function directly and do not require the lighting pair dataset.

1) Adversarial Relighting Network (ARNet): The proposed ARNet contains three modules, *i.e.*, *lighting prediction network (LPreNet)*, *adversarial lighting estimation network (AdvLNet)*, and *lighting rendering network (LRenNet)*. Given a face image $\mathbf{I}$, we first feed it to the LPreNet and get the original lighting parameters $\mathbf{L}$, the face embedding and light embedding, respectively, as the stage ① shown in Fig. 3. Then, we use AdvLNet to estimate the adversarial lighting based on the three outputs (See the stage ② in Fig. 3). Finally, we render the relighted face via the LRenNet by concatenating the embeddings of face and adversarial lighting as inputs, as the stage ③ shown in Fig. 3. For the LPreNet and LRenNet, we regard it as the hourglass network [19]. The AdvLNet serves as a transformation for the light embedding, mapping the original light to the adversarial counterpart. We construct the AdvLNet with three fully connection layers that are linked by two relu layers. Moreover, to let AdvLNet adapt to different face appearances, we propose to add a dynamic convolution layer whose convolutional weights are estimated by a fully connection layer and the face embedding. Specifically, as shown in Fig. 3, we first feed the deep feature of the original face into a fully connection layer and get a 1×81 vector. Then we reshape the vector to form a 9×9 matrix, which is just the dynamic convolutional weights. Next, we apply the 9×9 convolutional layer to the feature of the original lighting and finally generate the adversarial lighting. Notes, refer to Eqs. (1)–(3) we use the spherical harmonic form to indicate the adversarial lighting.

2) Loss Functions and Training Details: We train the three networks in a two-stage way. For the first stage, we train the LPreNet and LRenNet by regarding them as a pure deep relighting task, excluding the AdvLNet. Following the setups of [19], we use the CelebA-HQ dataset [45] and generate a training example by randomly selecting an original image $\mathbf{I}$ and a targeted relighting image $\mathbf{I}_t$ with their corresponding ground truth lighting $\mathbf{L}$ and $\mathbf{L}_t$, respectively. The LPreNet is fed with the original image $\mathbf{I}$ and predicts the lighting parameters $\mathbf{L}$; the LRenNet takes the targeted lighting parameters $\mathbf{L}_t$ and the face embedding of $\mathbf{I}$ as inputs and estimate the relighting counterpart $\mathbf{I}_t$. Under the supervision of the ground truths of relighted image and light, *i.e.*, $\mathbf{I}_t$ and $\mathbf{L}_t$, we can train LPreNet and LRenNet in an end-to-end strategy. Please refer to [19] for details.

After the first stage, we get the LPreNet and LRenNet for lighting prediction and rendering, respectively. For the second stage, we fix the two pre-trained networks, and tune the parameters of AdvLNet under the supervision of a FR method, *e.g.*, $\phi(\cdot)$, with the following loss function

$$\mathcal{L}_{\text{adv}} = \alpha \cos(\phi(\hat{\mathbf{I}}), \phi(\mathbf{I})) + \beta \frac{1}{N} \|\hat{\mathbf{I}} - \mathbf{I}\|_1, \tag{6}$$

where $\hat{\mathbf{I}} = \text{LRenNet}(\text{AdvLNet}(\text{LPreNet}(\mathbf{I})))$, N is the number of pixels, α and β are the weights. We set $\alpha = 1$ and $\beta = 1$ in the experiments. The first term is the same with Eq. (5) and makes sure the relighted face can fool the FR method. The second term is to limit the potential face variation after adversarial relighting. In the field of adversarial attack, it is desired to have small variation on the original image while misleading the targeted FR method.

In terms of the training details, we use the datasets of VGGFace2 [46] or CelebA [47] and employ the stochastic gradient descent (SGD) optimizer with learning rate 10^{-3} and momentum 0.9 to update AdvNet's parameters. We use mini-batch training strategy with batch size 8 and train AdvLNet with ten epochs. Please refer to Section IV-A for more details on datasets for training.

C. Physical ARA (Phy-ARA)

In addition to the digital lighting attack methods (*i.e.*, AQ-ARA and AP-ARA), we study a more important problem, *i.e.*, whether the adversarially relighted face could be reproduced in the real world. To this end, we propose the physical adversarial relighting attack (Phy-ARA). The main idea follows three steps: First, in a real-world scenario, we take a photo of a volunteer's face with a fixed camera under the natural light. Then, we use the AQ-ARA or AP-ARA to perform the attacks under the guidance of a face recognition (FR) method and get an adversarially relighted face $\hat{\mathbf{I}}$ and the adversarial lighting $\hat{\mathbf{L}}$. Finally, we can reproduce the adversarial lighting condition via a physical light source and take a new photo of the volunteer.

Although above process seems simple, one key issue makes it unavailable, that is, it is difficult to set suitable physical light sources meeting the pattern of estimated adversarial lighting. We address this problem from two aspects. First, we use the commonly-used point light source (PLS) to simulate the estimated adversarial lighting. Second, following the state-of-the-art active lighting recurrence (ALR) method [48], we can physically adjust the position of the PLS by a robotic arm and produce the real lighting condition that is the same to the estimated adversarial one. Fig. 4 shows the working scene and pipeline of Phy-ARA.

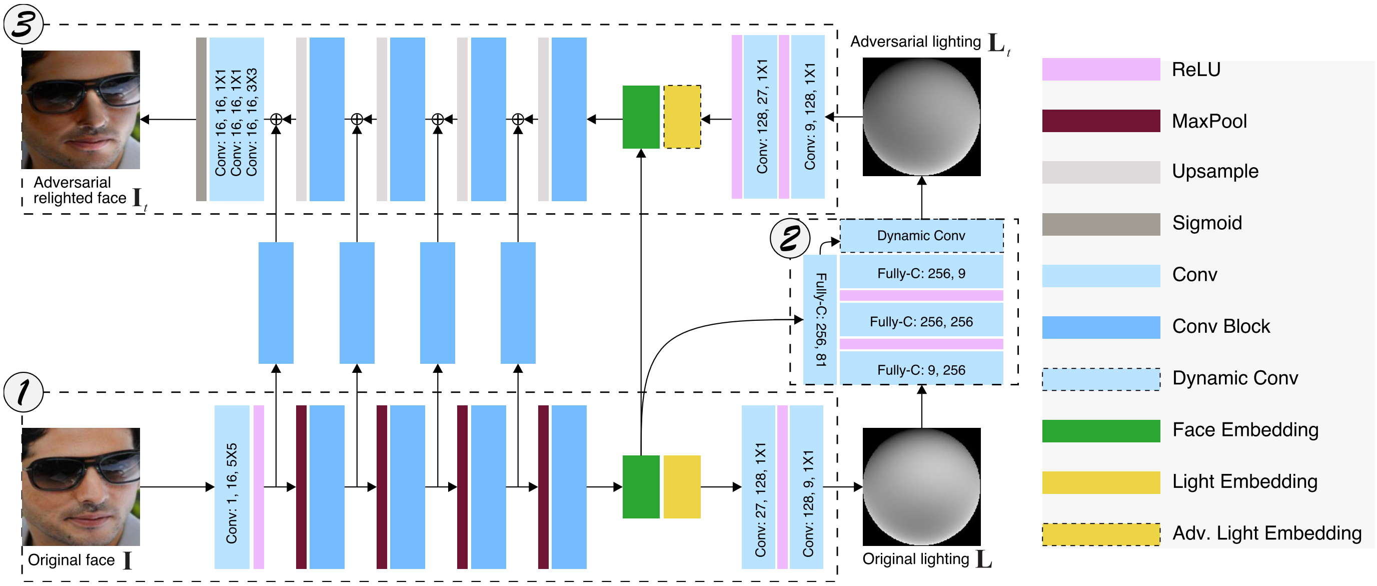


Fig. 3. Architecture of the proposed adversarial relighting network (ARNet). It contains three modules, *i.e.*, ① lighting prediction network (LPreNet), ② adversarial lighting estimation network (AdvLNet), and ③ lighting rendering network (LRenNet). The 'Conv Block' contains two convolution layers with the size of 3 × 3. The first convolutional layer is followed by a BatchNorm layer and a ReLU layer while the second one is only followed by a BatchNorm layer. The 'Dynamic Conv' means that the weights of the convolution layer is estimated from a fully connection layer, which makes the adversarial lighting adapt to different face embeddings.

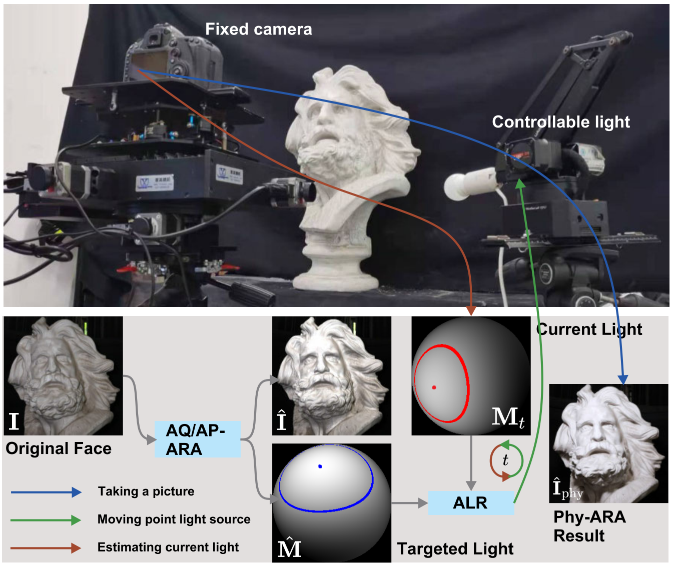


Fig. 4. Pipeline of Physical ARA. See text for details.

In contrast to existing ALR method [48] that depends on the parallel lighting model, our adversarial relighting is based on the spherical harmonic form that is a more general lighting representation and cannot be processed via the ALR directly. To fill the gap, we change the generation manner of lighting map in the ALR and enable it to support our experiment. Specifically, given the estimated adversarial lighting $\hat{\mathbf{L}}$, we generate the respective lighting map by $\hat{\mathbf{M}} = \mathrm{f}(\mathbf{N}^{\mathrm{s}}, \hat{\mathbf{L}})$, where $\mathbf{N}^{\mathrm{s}}$ indicates the normal of a sphere. After that, given the ith frame of current camera, *i.e.*, $\mathbf{I}_i$, the estimated scene normal $\mathbf{N}$ and the scene reflectance $\mathbf{R}$, we can calculate the corresponding spherical harmonic coefficients $\mathbf{L}_i$ by Eq. (1). Then, we calculate the lighting map of $\mathbf{I}_i$ by $\mathbf{M}_i = \mathrm{f}(\mathbf{N}^{\mathrm{s}}, \mathbf{L}_i)$ (See Fig. 4). The brightest position of $\mathbf{M}_i$ encodes the light source position in the azimuthal and polar axes, and the area of isointensity circle encodes the distance between light source and scene. Therefore, as shown in Fig. 4, we can get the instant navigation feedback from $\hat{\mathbf{M}}$ and $\mathbf{M}_i$. After that, we employ the incremental adjustment strategy in [48] to actively tune a robotic arm and update $\mathbf{L}_i$ to match $\hat{\mathbf{L}}$. Finally, we get the physical adversarial light (*i.e.*, $\hat{\mathbf{L}}_{\mathrm{phy}}$) under which we take a new image as the physical adversarial relighting image (*i.e.*, $\hat{\mathbf{I}}_{\mathrm{phy}}$). As shown in Fig. 4, we can clearly see that the image $\hat{\mathbf{I}}_{\mathrm{phy}}$ basically has the same lighting distribution as the $\hat{\mathbf{I}}$, showing effectiveness of our physical experiment.

IV. Experimental Results

A. Setups

1) Datasets: Adversarial relighting aims to fool face recognition method and relates to two tasks, *i.e.*, image relighting and face recognition. However, existing image relighting datasets do not contain the face identity annotations, thus are not suitable for evaluating our work. We conduct experiments on FR datasets, *i.e.*, VGGFace2 [46] and CelebA [47]. Both datasets can be employed to train AP-ARA and evaluate the performance of FR attacks. In addition, the informative attribute annotations in the CelebA dataset support the statistical analysis of the adversarial examples generated by FR attacks. For the attacking evaluation, we use testing datasets of VGGFace2 and CelebA, including 169k images of 500 identities and 19,962 images of 1,000 identities, respectively.

2) Attack Pipeline and Metrics: Let $\mathbf{I}$ be a face image. Given a face recognition model $\phi(\cdot)$, we can use the AQ-ARA (*i.e.*, solving Eq. (5) and Eq. (1)) or AP-ARA (*i.e.*, relying on a pre-trained network supervised by $\phi(\cdot)$) to relight $\mathbf{I}$ and produce a relighted face image $\hat{\mathbf{I}}$. After that, we can

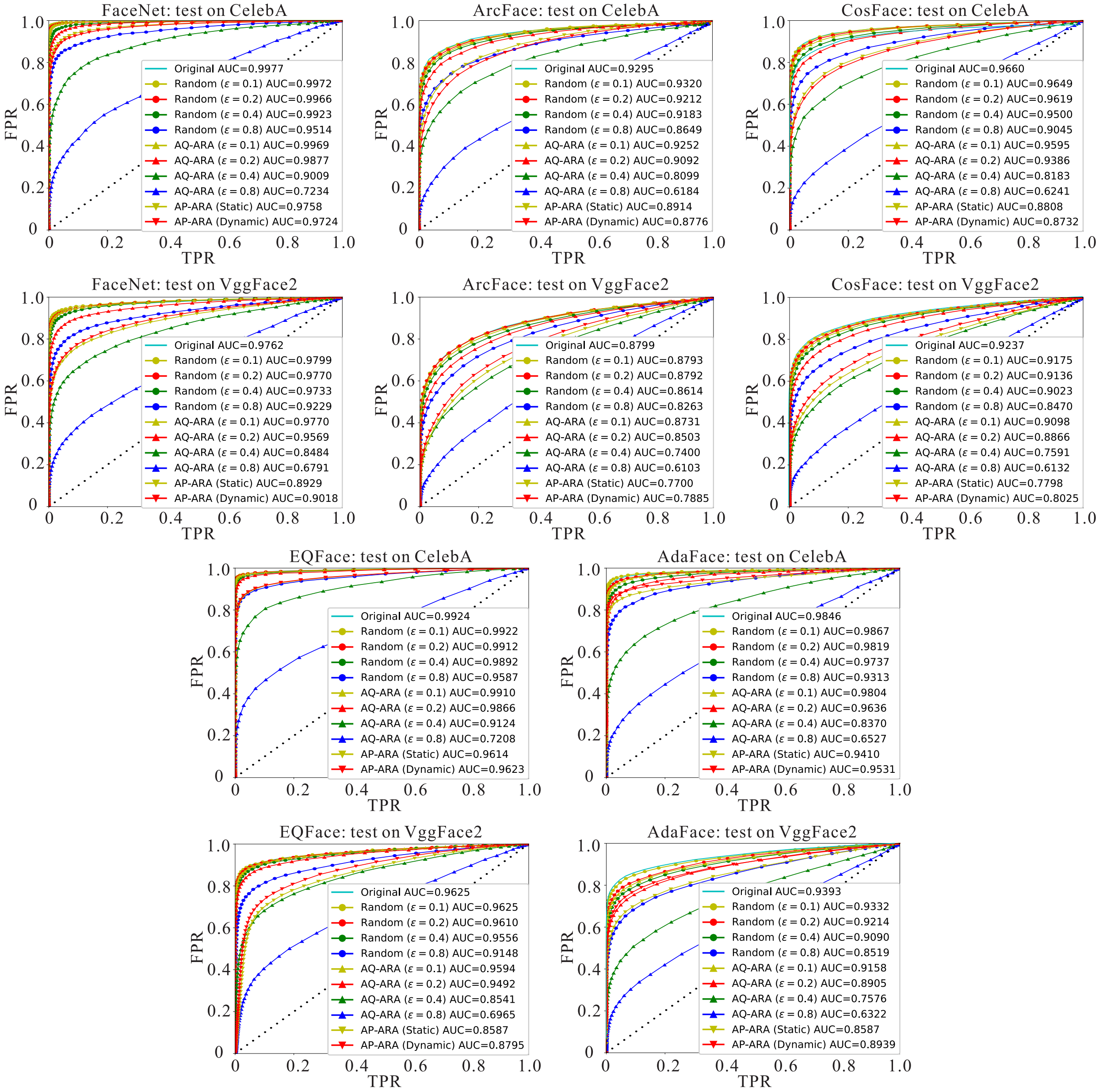


Fig. 5. ROC curves of various FR methods' performance under white-box attack. "AQ-ARA ($\epsilon = 0.1$)" indicates that the attack is launched with parameter $\epsilon = 0.1$.

evaluate the attack performance to see whether a FR model can still achieve high performance on the relighted faces. When the evaluated FR model is the same with the model to guide or supervise AQ-ARA or AP-ARA, we mean it the white-box attack (See Section IV-B), otherwise we denote it as transfer-based attack (See Section IV-C).

In terms of evaluation process, given a dataset for face recognition, we have n identities and $2k$ face images per identity. For the ith identity, we separate the $2k$ face images into two subsets, *i.e.*, the reference subset $\mathcal{R}_i$ and the targeted subset $\mathcal{T}_i$. Then, we use our ARAs and baselines to relight the face images in $\mathcal{T}_i$ and get adversarially relighted examples $\mathcal{A}_i$. All the reference subsets $\mathcal{R}_i$ consist of a larger set $\mathcal{R} = \cup_{i=1}^{n} \mathcal{R}_i$ while all subsets $\mathcal{A}_i$ make up of another set $\mathcal{A} = \cup_{i=1}^{n} \mathcal{A}_i$ After that, we calculate the cosine similarity between face images from $\mathcal{R}$ and $\mathcal{A}$ based on a FR model. As a result, we can get a $nk \times nk$ matrix $\mathbf{S}$. A robust FR model is desired to have a $\mathbf{S}$ with the block-diagonal pattern, that is, the face images from the same identify should have high cosine similarity tending to 1 otherwise having low similarities tending to -1. We can also define a ground-truth matrix $\mathbf{G} \in \mathbb{R}^{nk \times nk}$ for face recognition where $G(i, j) = 1$ means the ith face in $\mathcal{R}$ and jth face in $\mathcal{A}$ are from the same identity otherwise $G(i, j) = 0$. We can calculate the true positive rate (TPR) and false positive rate (FPR) by comparing $\mathbf{S}$ with $\mathbf{G}$ under a given cosine similarity threshold. With a series of thresholds, we can draw the ROC curve and its area under curve (AUC) to measure the effectiveness of attacks, that is, a more effective adversarial relighting attack corresponds to a lower AUC. In this paper, we select sixteen face images per identity, *i.e.*, $k = 8$.

In this paper, we use two limited vision image quality assessment metrics, *i.e.*, BRISQUE [49] and NIQE [50], to compare the image quality of generated adversarial exam-

TABLE I

AVERAGE AUC VALUES AND IMAGE QUALITY SCORES OF RANDOM RELIGHTING, AQ-ARA AND AP-ARA FOR ALL THE FR METHODS ON CELEBA AND VGGFACE2 DATASETS

Datasets		CelebA			VGGFace2		
Metrics		AUC↓	BRISQUE↓	NIQE↓	AUC↓	BRISQUE↓	NIQE↓
Original		0.9797	33.52	4.25	0.9623	41.75	5.12
Random	$\epsilon = 0.1$	0.9791	33.46	4.25	0.9610	41.62	5.01
	$\epsilon = 0.2$	0.9723	33.46	4.23	0.9604	41.56	5.02
	$\epsilon = 0.4$	0.9704	34.22	4.36	0.9601	42.64	5.31
	$\epsilon = 0.8$	0.9418	38.14	4.98	0.9023	43.89	5.52
AQ-ARA	$\epsilon = 0.1$	0.9782	33.34	4.19	0.9616	42.35	4.99
	$\epsilon = 0.2$	0.9701	34.58	4.39	0.9588	41.68	4.97
	$\epsilon = 0.4$	0.9124	39.97	4.93	0.8593	45.35	5.42
	$\epsilon = 0.8$	0.7094	44.68	5.73	0.6603	47.51	6.13
AP-ARA	Static	0.9538	36.68	4.33	0.8752	41.22	4.92
	Dynamic	0.9529	36.60	4.32	0.8729	43.24	5.19

ples. A larger BRISQUE or NIQE score indicates worse quality and less naturalness of an image.

3) Models: We evaluate the proposed attacks and baselines against several face recognition models, *i.e.*, FaceNet [51], CosFace [52], ArcFace [53], EQFace [54], and AdaFace [55]. All these models require face images to be pre-processed by MTCNN [56] for best performance, and we will do so as usual.

4) Baseline Methods: Our adversarial relighting only tunes the light parameters (*i.e.*, nine spherical harmonic coefficients) to fool face recognition system, leading to smooth variation of the face image. In contrast, existing additive-perturbation-based adversarial attacks can tune each pixel independently under the guidance of FR models [31], [32], [33]. For the fairness of the comparison, we do not include these attacks as part of the baselines. Actually, a reasonable baseline is to conduct random relighting without any FR model guidance. To this end, we can first randomly sample the nine lighting coefficient variations within the range $[-\epsilon, \epsilon]$ and use relighting methods to apply the sampled lights to the targeted face. We employ deep portrait relighting (DPR) [19] since it achieves the state-of-the-art relighting performance and show more realistic results than other methods, *e.g.*, SfSNet [57]. To verify the effectiveness of AP-ARA, we perform cross-validation between the testing sets of VGGFace2 and CelebA. Specifically, we train our network (AP-ARA) on the testing set of VGGFace2 and test on the testing set of CelebA. Similarly, we can also train our network on the testing set of CelebA and test on the testing set of VGGFace2. As for the random manner and AQ-ARA, they have no training process and just select or estimate a target lighting to produce the relighted image. Thus, we can directly test the two methods on VGGFace2 and CelebA without cross validation. All experiments are conducted on a computer with an NVIDIA RTX 2080 GPU. We set the term weights $\alpha = 1$ and $\beta = 1$ (see Eq. (6)) for our AP-ARA.

B. Comparison Results on White-Box Attack

We conduct different lighting attacks against SOTA face recognition (FR) methods on two datasets, *i.e.*, CelebA and VGGFace2. The SOTA FR methods include FaceNet [51], CosFace [52], ArcFace [53], EQFace [54], and AdaFace [55]. The lighting attack manners include random relighting, the proposed AQ-ARA and AP-ARA. For the random relighting and the proposed AQ-ARA, we set the parameter $\epsilon \in \{0.1, 0.2, 0.4, 0.8\}$. The proposed AP-ARA is trained on CelebA and VGGFace2 datasets, respectively. The results of the above experiments are shown in Fig. 5 and Table I. We have the following observations: ❶ As the ϵ becomes larger, the AUC under AQ-ARA gradually reduce with slight increasing of the BRISQUE and NIQE, demonstrating the effectiveness of the objective function Eq. (5). ❷ When comparing the random relighting baseline with AQ-ARA, we see that AQ-ARA leads to lower AUC under the same ϵ, demonstrating that AQ-ARA can is able to find adversarial light that misleads FaceNet easily. ❸ AP-ARA methods' AUCs are between the results of AQ-ARA with $\epsilon = 0.2$ and $\epsilon = 0.4$, but usually hold a better image quality (lower BRISQUE and NIQE) than AQ-ARA ($\epsilon = 0.4$). In addition, the computational cost of AP-ARA is lower than that of AQ-ARA, due to the one-step ARNet in the test procedure. ❹ When comparing the two variants of AP-ARA, AP-ARA-Dynamic gets lower AUC than AP-ARA-Static on VGGFace2 while having similar results on CelebA, hinting the effectiveness of the proposed ARNet that is able to adapt to different face via a dynamic convolution.

We also provide the two visualization results in Fig. 6 and have the following observations: ❶ Both AQ-ARA and AP-ARA are able to produce the realistic relighting pattern. More specifically, AQ-ARA tends to generate the adversarial light along a specific direction. For example, in second case, the adversarial light of AQ-ARA is along the northwest-southeast direction. The first case has similar result with a different direction. In contrast, AP-ARA tends to generate more complex light pattern that cannot be regarded as a directional light. ❷ When comparing the random relighting with AQ-ARA, we see that AQ-ARA always lets the FaceNet generate smaller cosine similarity under the same ϵ, which further demonstrates the effectiveness of the proposed method. ❸ When comparing AP-ARA-Static with AP-ARA-Dynamic, the relighted faces of AP-ARA-Dynamic have lower similarity to the reference one.

C. Comparison Results on Transferability

We then verify the transferability of the adversarial relighting. Specifically, we employ three FR methods, i.e., FaceNet [51], ArcFace [53] and CosFace [52]. Next, we use the adversarial relighting examples crafted from one FR method to attack another FR method. Notes, since the random manner and AQ-ARA have no training process and just select or estimate a target lighting to produce the relighted image, only the AP-ARA has the transferability experiments. The above experiment results are shown in Fig. 7. The results show significant transferability over the two datasets for all these FR methods. It means that our AP-ARA method can be used to black-attack for current FR methods, and it can achieve comparable performance to the white-box attack results.

In the above experiments, we set $\alpha = 1$ and $\beta = 1$ in the loss function, i.e., Eq. (6). We further explore the

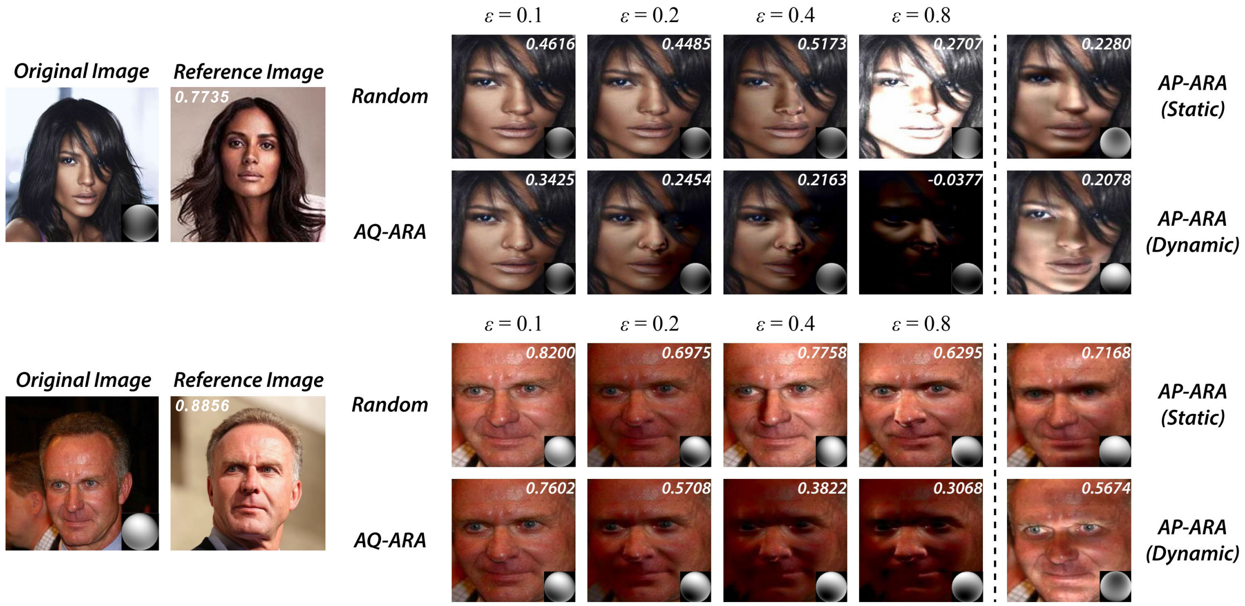


Fig. 6. Visualized adversarial examples of the proposed attacks and baselines. We vary the ϵ parameter to show the relighted faces and also place the shading sphere in the lower right corner of each face image.

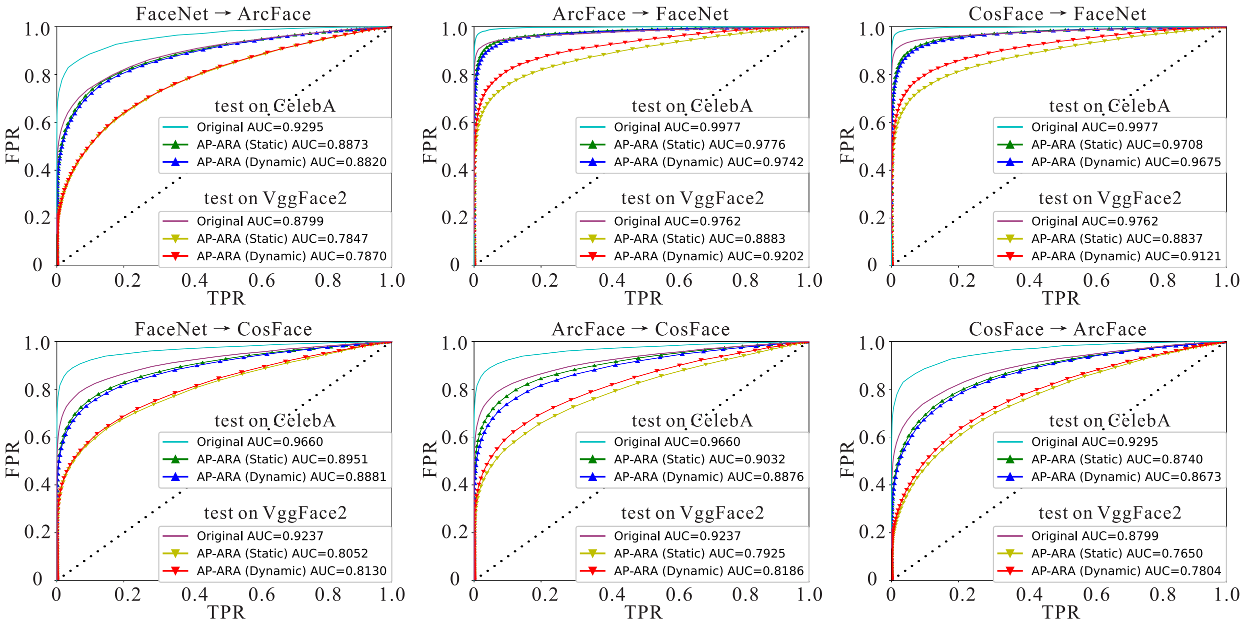


Fig. 7. ROC curves of various FR methods' performance under black-box attack. "A → B" means training by A and testing by B.

influence of different weight settings to our AP-ARA. We train FaceNet [51] on CelebA [47] dataset and test it on VggFace2 dataset under different weight settings of Eq. (6). All the comparison results are shown in Fig. 8. The first term of Eq. (6) makes sure the relighted face can fool the FR method and the second term is to limit the potential face variation and encourages the image authenticity after adversarial relighting. As shown in Fig 8, when α is larger relative to β, although the lighting attack effect is stronger (lower AUC score), the authenticity of the image is also worse (refer to the cases in the right part of Fig. 8). On the contrary, when β is larger relative to α, although the image authenticity is better, the lighting attack effect is weaker (larger AUC score). In the field of adversarial attack, it is desired to guarantee naturalness and authenticity of the relighted image while misleading the targeted FR method. Therefore, taking the above two factors into consideration, we finally choose $\alpha = 1$ and $\beta = 1$ in Eq. (6).

D. Analyzing Effects of Light to Face Recognition via AQ-ARA

In this section, we study the effects of light to FR via the proposed adversarial attack. Specifically, for the ith example, we have a pair of adversarial light $\hat{\mathbf{L}}_i$ and the original light $\mathbf{L}_i$ via the AQ-ARA, and we denote their lighting maps as $\mathbf{M}_i$ and $\hat{\mathbf{M}}_i$, respectively. Then, we can calculate the difference between lighting maps of $\hat{\mathbf{L}}_i$ and $\mathbf{L}_i$ by $\mathbf{D}_i = |\hat{\mathbf{M}}_i - \mathbf{M}_i|$. and

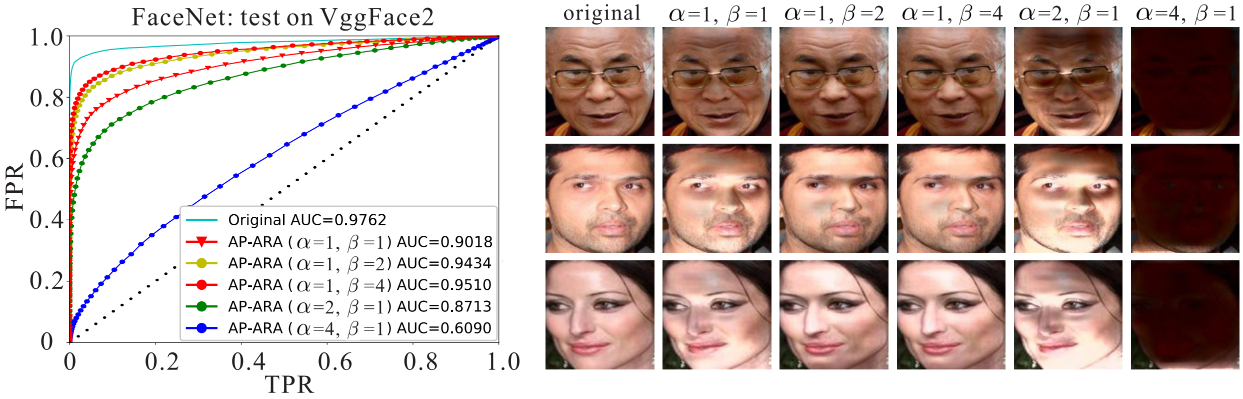


Fig. 8. Comparison results of varied weight settings in the loss function, i.e., Eq (6).

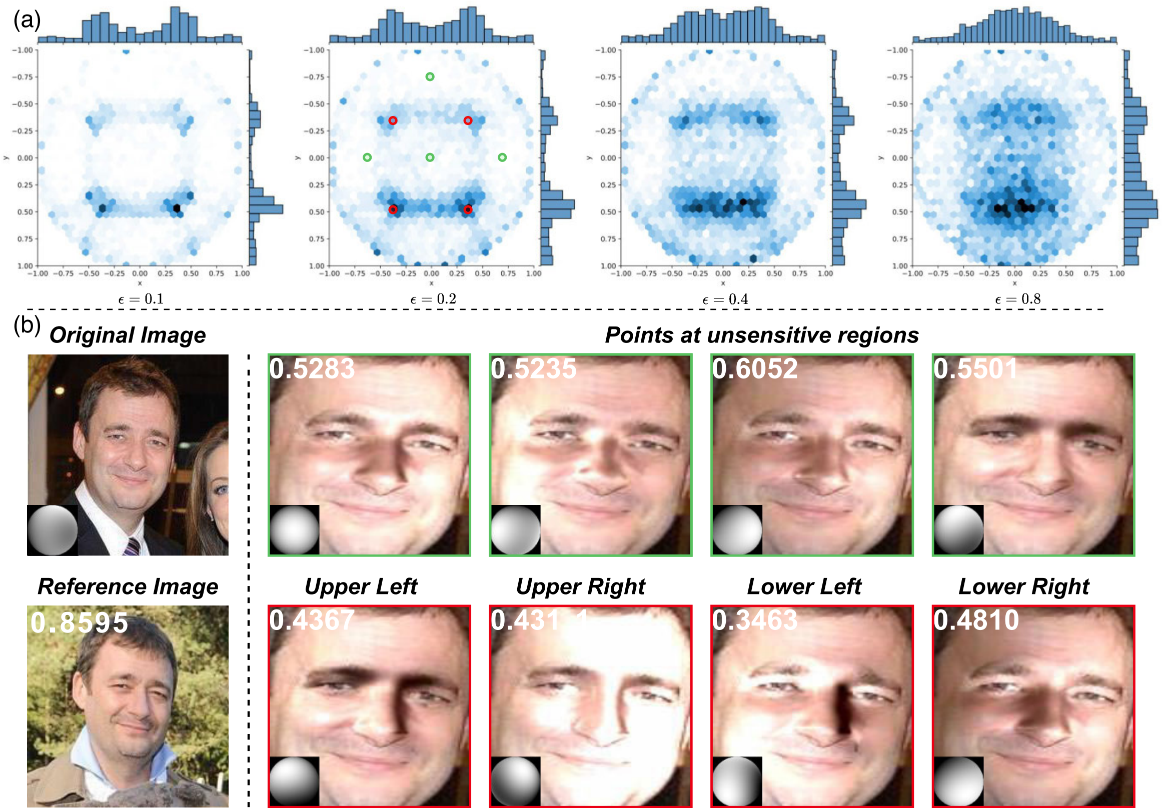


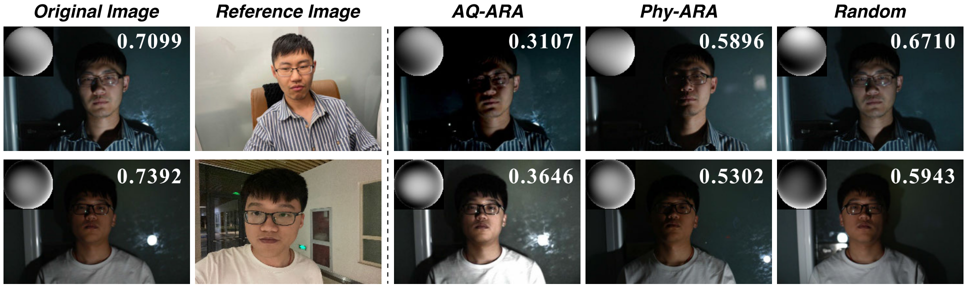
Fig. 9. (a) Four histogram maps under $\epsilon \in \{0.1, 0.2, 0.4, 0.8\}$. Please find detailed descriptions in the text. (b) Relighting examples based on the light conditions defined in (a)-'$\epsilon = 0.2$'. The similarity of all examples to the reference image are labeled at the upper-left corner.

get the maximum difference position (*i.e.*, sensitive points) on the map by $(x_i, y_i) = \arg\max_{(x,y)} \mathbf{D}_i$. After that, we can calculate a 3D histogram on hexagonal grids by counting the number of sensitive points in each grid. As a result, we can get four histogram maps under four ϵ and show them in Fig. 9 (a). The higher values in the figure correspond to sensitive lighting sources that may fool face recognition easily. We also provide relighting examples based on the sensitive and insensitive lighting conditions.

We have following observations: ❶ According to the difference maps under $\epsilon = 0.1$ and $\epsilon = 0.2$, the main differences locate at the four positions around the center, indicating the sensitive lighting sources to face recognition. For example, in Fig. 9 (b), the sensitive lights indicated by red points lead to smaller similarity to the reference image. ❷ As the ϵ becomes larger, the sensitive lighting sources increase because the larger ϵ allows larger variation of faces. Previous less sensitive lighting sources are also able to fool FR methods. ❸ According to the results of $\epsilon = 0.4$ and $\epsilon = 0.8$, FR method is more sensitive to the lighting sources at the bottom.

E. Comparison With Other Adversarial Attack Methods

We also compare our adversarial relighting method with other adversarial attack types, including adversarial noise, adversarial patch and adversarial make-up. Adversarial noise, adversarial patch and adversarial relighting (this paper) belong to dodging attack, which aims to reduce the confidence of the



Average scores under 10 subjects

Original Image	AQ-ARA	Phy-ARA	Random
0.7492	0.3245	0.5603	0.6133

Fig. 10. The cosine similarity scores between the reference image and the original image, AQ-ARA image, Phy-ARA image, random relighting image, respectively. The lower the score, the better the attack performance. The Phy-ARA can achieve better attack performance than the random one and generate more realistic image than AQ-ARA.

TABLE II

Comparison Results of Our Method With Different Adversarial Attack Types Against FaceNet [51] on CelebA Dataset [47], Including Adversarial Noise, Adversarial Patch and Adversarial Make-up Methods. Notes, All the ASR-Imp Scores Come From [59] and All the ASR-Dod Scores Come From [64]

Attack Type	Method		ASR-dod↑ (Eq. 14 in [58])	ASR-imp↑ (Eq. 12 in [59])
Noise	FGSM [60]		–	1.9
	PGD [61]		–	3.5
	MI-FGSM [33]		0.9900	4.6
Patch	Adv-Hat [62]		–	4.7
	Adv-Glasses [63]		–	9.1
	Gen-AP [64]		0.9975	15.8
Make-up	Semantic-Adv [65]		–	9.0
	Adv-Face [66]		–	21.6
	Adv-Attribute [59]		–	31.8
Lighting	AP-ARA	Static	0.9982	14.9
		Dynamic	0.9988	15.5

similarity of the same identity pair in order to evade face recognition. Adversarial make-up belongs to impersonation attack, which aims to cause the input face image to be identified as any other individual in the face database. We use attack success rates (ASR) to evaluate the performance of these adversarial methods. Notes, dodging attack takes images of the same identity as input and impersonation attack takes images of different identities for evaluation. These two kinds of attacks have different approaches to calculate attack success rates. In this paper, we use "ASR-dod" and "ASR-imp" to denote the attack success rates of dodging attack and impersonation attack. Let $< \mathbf{I}, \mathbf{I}^{\text{dod}} >$ and $< \mathbf{I}, \mathbf{I}^{\text{imp}} >$ be the image pair from the same identity and different identities, respectively. $\hat{\mathbf{I}}$ is the adversarial attack image of $\mathbf{I}$. Refer to Eq. 14 in [58] and Eq. 12 in [59], the formulas of ASR-dod and ASR-imp are shown as follows:

$$\text{ASR-dod} = \frac{1}{P}\sum_{p}^{P} 1_{\tau}[||\phi(\hat{\mathbf{I}}), \phi(\mathbf{I}^{\text{dod}})||_2 < \tau^{\text{dod}}], \tag{7}$$

$$\text{ASR-imp} = \frac{1}{P}\sum_{p}^{P} 1_{\tau}[\cos(\phi(\hat{\mathbf{I}}), \phi(\mathbf{I}^{\text{imp}})) > \tau^{\text{imp}}] \times 100, \tag{8}$$

where P is the number of image pairs, 1_{τ} denotes the indicator function, $\phi(\cdot)$ indicates a face recognition network generating the face embedding. Following the same parameter setting of baselines, we set $\tau^{\text{dod}} = 0.4$ and $\tau^{\text{imp}} = 0.6$.

The comparison results are shown in Table II. The arrow ↑ means that the higher the score is, the stronger the attack of the method. First, we can see that our adversarial relighting method (AP-ARA) can achieves better performance (higher ASR-dod scores) than adversarial noise and adversarial patch methods. The reason is that the lighting attack is global and may be easier to fool the FR system than other attack types. Second, we can see that the ASR-imp scores of our AP-ARA cannot overcome the adversarial make-up methods. In fact, the impersonation attack (adversarial make-up) aims to cause the input face image to be identified as other individual and the dodging attack (adversarial noise, adversarial patch and our adcversarial relighting) lacks of relevant supervision during training. Thus, it is perfectly normal that our method achieves lower ASR-imp scores (both for the adversarial noise and adversarial patch methods) than adversarial make-up methods.

F. Validating Physical Attack via Phy-ARA

We follow the steps of Section III-C to validate the proposed Phy-ARA through a volunteer. Specifically, we first take an image of the volunteer with a natural light source as the original image. Then, we conduct the AQ-ARA and produce a relighted face and the adversarial light. After that, we drive a bulb fixed at a robotic arm to fit the adversarial light. Finally, we take a new picture as the result of Phy-ARA and calculate the cosine similarity between the reference image and other images based on the FaceNet. The lower the score, the better the attack performance. We conduct the above experiment on 10 subjects, the avearges cosine similarity scores and two visual cases are shown in Fig. 10. We can see that: ❶ Comparing the cosine similarity scores of the original and relighted images, we find that the adversarial light produced by AQ-ARA indeed affects FaceNet significantly, which reduces the similarity from 0.7 to 0.3. The respective physical counterpart reduces the similarity to 0.5. In contrast, the random relighting affects the face recognition slightly.

TABLE III

ASR-Dod (Attack Success Rate, i.e., Recognition Failure Rates) Sorces for Different Models, Including the Original FaceNet, Finetuned FaceNet by Our AP-ARA With 10% and 30% Lighting Attacks.The Lower the ASR-Dod Score Is, the Better the Performance of FaceNet. As the Proportion of Samples With Lighting Attack in the Training Set Increases, the Lighting Robustness of FaceNet Has Been Improved (From 0.9988 to 0.9797), but the Performance of FaceNet Slightly Deteriorates (From 0.0245 to 0.0309) When Testing on the Original (Without Lighting Attack) Testing Set. In a Word, Improving the Lighting Robustness While Maintaining the Original Performance of Current FR System Is Quite a Challenging Task, Which Warrants Further Investigation and Exploration

model	test on original testing set	test on testing set with lighting attack
original FaceNet	0.0245	0.9988
finetune with 10% lighting attack	0.0305	0.9855
finetune with 30% lighting attack	0.0309	0.9797

❷ The Phy-ARA can achieve better attack performance than the random one and generate more realistic image than AQ-ARA. It means that our method can be effectively used in real-world lighting attack task.

The objective of realistically reproducing lighting is to physically place/move the light source within a scene to recur the given target lighting condition as closely as possible. Realistically reproducing lighting remains a very challenging problem. To the best of our knowledge, active lighting recurrence (ALR) [48] is currently the only method available for realistically reproducing lighting. ALR assumes that the illumination comes from only a light source, and thus accurate reproduction of lighting condition can only be achieved when the target lighting also originates from a light source. Given a target global illumination (such as represented by spherical harmonics), how many light sources should be selected to achieve the optimal approximate effect? Which position should the light source be moved to that best approximates the target lighting? Which type (e.g., point light source, surface light source, *etc.*) of light source will be better? These issues require further exploration and research.

G. Improving Lighting Robustness of Current FR method

This paper proposes a new adversarial attack manner, i.e., adversarial relighting. In the above experiments, we have verified that our adversarial relighting method can effectively fool current face recognition systems and reduce the recognition rate of them. In fact, our method can also be used to improve the lighting robustness of current FR method. To verify this, we finetune FaceNet with the help of our AP-ARA. Sepcifically, for an original training sample that including a face image pair (e.g., $< \mathbf{I}, \mathbf{I}' >$) and the label (1 or 0, indicates the two images are from the same identity or not), we generate the lighting attack image $\hat{\mathbf{I}}$ of $\mathbf{I}$ by our AP-ARA, and construct the new training sample (including image pair $< \hat{\mathbf{I}}, \mathbf{I}' >$ and the label), i.e., the relighted training sample, which can be used to improve the lighting robustness of current FR methods by finetuning. In order to balance the performance of FR method to the lighting attack condition and the original condition (without lighting attack), we construct relighted training samples according to a certain proportion (i.e., 10% or 30% in Table III) for all the original training samples. For the original testing set, we also construct the testing set with lighting attack. The FaceNet is trained by using 100 epochs. Learning rate is initially set to 0.00001 and then divided by 2 after finishing 20 epochs.

Table III shows the ASR-dod (attack success rate, i.e., recognition failure rates) sorces for different models, including the original FaceNet, finetuned FaceNet by our AP-ARA with 10% and 30% lighting attacks. Refer to Eq. (7) for the formula of ASR-dod. The lower the ASR-dod score is, the better the performance of FaceNet. As the proportion of samples with lighting attack in the training set increases, the lighting robustness of FaceNet has been improved (from 0.9988 to 0.9797), but the performance of FaceNet slightly deteriorates (from 0.0245 to 0.0309) when testing on the original (without lighting attack) testing set. In a word, improving the lighting robustness while maintaining the original performance of current FR system is quite a challenging task, which warrants further investigation and exploration.

V. Conclusion

In this work, we have unveiled a new adversarial threat for FRS from the face lighting perspective and investigated a new task, the adversarial relighting attack (ARA). The ARA aims to produce a high-realism relighted face image, which can able to fool the SOTA deep face recognition (FR) methods. We first designed the physical model-based ARA denoted as albedo-quotient-based adversarial relighting attack (AQ-ARA), which can generate natural adversarial light and synthesize adversarially relighted face images. To better suit efficiency-sensitive applications, we further proposed the auto-predictive adversarial relighting attack (AP-ARA) by training an adversarial relighting network to automatically predict the adversarial light in a one-step manner according to different input faces. More importantly, through a precise relighting device, we are able to transfer the above digital adversarial attacks to physical ARA (Phy-ARA), making the estimated adversarial lighting condition reproducible in the real world. The extensive evaluation of the proposed method on various SOTA FRS has demonstrated the feasibility of generating adversarially relighted faces to fool the FRS with ease. When the input faces are under other natural degradations such as very low-resolution, heavy occlusion, *etc.*, we expect that the proposed ARA would not be as effective. How to robustify the proposed ARA warrants a future study. Bad actors can potentially capitalize on the proposed ARA to fool some safety-critical FRS that are not yet prepared for this new type of attack. We hope that this work can also accelerate the R&D of next-generation adversarially robust FRS. Moreover, we can use the proposed adversarial righting attack to analyze other face-related tasks, *e.g.*, DeepFake detection [67], visual-based heart rhythm estimation [68], and facial age estimation [69].

References

[1] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov. 2015.

[2] W. Hu and H. Hu, "Disentangled spectrum variations networks for NIR–VIS face recognition," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1234–1248, May 2020.

[3] W. W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 327–340, Jan. 2012.

[4] R. Abiantun, F. Juefei-Xu, U. Prabhu, and M. Savvides, "SSR2: Sparse signal recovery for single-image super-resolution on faces with extreme low resolutions," *Pattern Recognit.*, vol. 90, pp. 308–324, Jun. 2019.

[5] F. Juefei-Xu, E. Verma, P. Goel, A. Cherodian, and M. Savvides, "DeepGender: Occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 136–145.

[6] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3539–3545.

[7] F. Juefei-Xu, K. Luu, and M. Savvides, "Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4780–4795, Dec. 2015.

[8] D. K. Pal, F. Juefei-Xu, and M. Savvides, "Discriminative invariant kernel features: A bells-and-whistles-free approach to unsupervised face recognition and pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5590–5599.

[9] B. Huang et al., "When face recognition meets occlusion: A new benchmark," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4240–4244.

[10] F. Juefei-Xu, D. K. Pal, and M. Savvides, "Hallucinating the full face from the periocular region via dimensionally weighted K-SVD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 1–8.

[11] F. Juefei-Xu and M. Savvides, "Fastfood dictionary learning for periocular-based full face hallucination," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2016, pp. 1–8.

[12] H. Fang, W. Deng, Y. Zhong, and J. Hu, "Generate to adapt: Resolution adaption network for surveillance face recognition," in *Proc. ECCV*, 2020, pp. 741–758.

[13] J. Liang, Y. Cao, C. Zhang, S. Chang, K. Bai, and Z. Xu, "Additive adversarial learning for unbiased authentication," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11420–11429.

[14] G. Srivastava and S. Bag, "Modern-day marketing concepts based on face recognition and neuro-marketing: A review and future research directions," *Benchmarking, Int. J.*, vol. 31, no. 2, pp. 410–438, Feb. 2024.

[15] T. Sun et al., "Single image portrait relighting," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Aug. 2019.

[16] Z. Wang, X. Yu, M. Lu, Q. Wang, C. Qian, and F. Xu, "Single image portrait relighting via explicit multiple reflectance channel modeling," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–13, Nov. 2020.

[17] A. Stoschek, "Image-based re-rendering of faces for continuous pose and illumination directions," in *Proc. IEEE CVPR*, vol. 1, 2000, pp. 582–587.

[18] X. Zhang, Y. Song, Z. Li, and J. Jiang, "PR-RL: Portrait relighting via deep reinforcement learning," *IEEE Trans. Multimedia*, vol. 24, pp. 3240–3255, 2022.

[19] H. Zhou, S. Hadap, K. Sunkavalli, and D. Jacobs, "Deep single-image portrait relighting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7193–7201.

[20] R. Yi, C. Zhu, and K. Xu, "Weakly-supervised single-view image relighting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8402–8411.

[21] Y. Liu, A. Neophytou, S. Sengupta, and E. Sommerlade, "Relighting images in the wild with a self-supervised Siamese auto-encoder," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 32–40.

[22] A. Shashua and T. Riklin-Raviv, "The quotient image: Class-based re-rendering and recognition with varying illuminations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 129–139, 2001.

[23] A. Hou, Z. Zhang, M. Sarkis, N. Bi, Y. Tong, and X. Liu, "Towards high fidelity face relighting with realistic shadows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14714–14723.

[24] R. Singh, A. Agarwal, M. Singh, S. Nagpal, and M. Vatsa, "On the robustness of face recognition algorithms against attacks and bias," in *Proc. AAAI*, vol. 34, no. 9, 2020, pp. 13583–13589.

[25] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, "Unravelling robustness of deep learning based face recognition against adversarial attacks," in *Proc. AAAI*, 2018.

[26] M. Ferrara, A. Franco, and D. Maltoni, "The magic passport," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep. 2014, pp. 1–7.

[27] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, "SWAPPED! Digital face presentation attack detection via weighted local magnitude pattern," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 659–665.

[28] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer, "Detecting facial retouching using supervised deep learning," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 9, pp. 1903–1913, Sep. 2016.

[29] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," 2018, *arXiv:1811.00656*.

[30] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.

[31] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICML*, 2015.

[32] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. ICLR*, 2017, pp. 99–112.

[33] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

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

[35] L. Gao, Z. Huang, J. Song, Y. Yang, and H. T. Shen, "Push & pull: Transferable adversarial examples with attentive attack," *IEEE Trans. Multimedia*, vol. 24, pp. 2329–2338, 2022.

[36] R. Sanchez-Matilla, C. Y. Li, A. S. Shamsabadi, R. Mazzon, and A. Cavallaro, "Exploiting vulnerabilities of deep neural networks for privacy protection," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1862–1873, Jul. 2020.

[37] Y. Cheng et al., "Pasadena: Perceptually aware and stealthy adversarial denoise attack," *IEEE Trans. Multimedia*, vol. 24, pp. 3807–3822, 2022.

[38] A. Bhattad, M. J. Chong, K. Liang, B. Li, and D. A. Forsyth, "Unrestricted adversarial examples via semantic manipulation," in *Proc. ICLR*, 2020.

[39] Q. Guo et al., "Watch out! Motion is blurring the vision of your deep neural networks," in *Proc. NIPS*, vol. 33, 2020, pp. 975–985.

[40] X. Jia, X. Wei, X. Cao, and X. Han, "Adv-watermark: A novel watermark perturbation for adversarial examples," in *Proc. ACM MM*, 2020, pp. 1579–1587.

[41] L. Zhai et al., "It's raining cats or dogs? Adversarial rain attack on DNN perception," 2020, *arXiv:2009.09205*.

[42] R. Gao, Q. Guo, F. Juefei-Xu, H. Yu, and W. Feng, "AdvHaze: Adversarial haze attack," 2021, *arXiv:2104.13673*.

[43] A. Shashua and T. Riklin-Raviv, "The quotient image: Class-based re-rendering and recognition with varying illuminations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 129–139, 2001.

[44] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 78–92, Jan. 2019.

[45] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. ICLR*, 2018.

[46] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

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

[48] Q. Zhang, W. Feng, L. Wan, F.-P. Tian, and P. Tan, "Active recurrence of lighting condition for fine-grained change detection," in *Proc. IJCAI*, 2018, pp. 4972–4978.

[49] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[50] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Apr. 2012.

[51] 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.

[52] H. Wang et al., "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[53] 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.

[54] R. Liu and W. Tan, "EQFace: A simple explicit quality network for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1482–1490.

[55] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18750–18759.

[56] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[57] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "SfSNet: Learning shape, refectance and illuminance of faces in the wild," in *Proc. IEEE CVPR)*, 2018, pp. 6296–6305.

[58] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1452–1466, 2021.

[59] S. Jia et al., "Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 34136–34147.

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

[61] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[62] 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.

[63] 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.

[64] Z. Xiao et al., "Improving transferability of adversarial patches on face recognition with generative models," in *IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2021, pp. 11845–11854.

[65] H. Qiu, C. Xiao, L. Yang, X. Yan, H. Lee, and B. Li, "SemanticAdv: Generating adversarial examples via attribute-conditioned image editing," in *Proc. ECCV*, 2020, pp. 19–37.

[66] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces: Adversarial face synthesis," 2019, *arXiv:1908.05008*.

[67] H. Qi et al., "DeepRhythm: Exposing DeepFakes with attentional visual heartbeat rhythms," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020.

[68] X. Niu, S. Shan, H. Han, and X. Chen, "RhythmNet: End-to-end heart rate estimation from face via spatial–temporal representation," *IEEE Trans. Image Process.*, vol. 29, pp. 2409–2423, 2020.

[69] S. Feng, C. Lang, J. Feng, T. Wang, and J. Luo, "Human facial age estimation by cost-sensitive label ranking and trace norm regularization," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 136–148, Jan. 2017.