Effective and Robust Physical-World Attacks on Deep Learning Face Recognition Systems

Meng Shen, *Member, IEEE*, Hao Yu, Liehuang Zhu, *Member, IEEE*, Ke Xu, *Senior Member, IEEE*, Qi Li, *Senior Member, IEEE*, and Jiankun Hu, *Senior Member, IEEE*

***Abstract*— Deep neural networks (DNNs) have been increasingly used in face recognition (FR) systems. Recent studies, however, show that DNNs are vulnerable to adversarial examples, which potentially mislead DNN-based FR systems in the physical world. Existing attacks either generate perturbations working merely in the digital world, or rely on customized equipment to generate perturbations that are not robust in the ever-changing physical environment. In this paper, we propose FaceAdv, a physical-world attack that crafts adversarial stickers to deceive FR systems. It mainly consists of a sticker generator and a convertor, where the former can craft several stickers with different shapes while the latter aims to digitally attach stickers to human faces and provide feedback to the generator to improve the effectiveness. We conduct extensive experiments to evaluate the effectiveness of FaceAdv on attacking three typical FR systems (i.e., ArcFace, CosFace and FaceNet). The results show that compared with a state-of-the-art attack, FaceAdv can significantly improve the success rates of both dodging and impersonating attacks. We also conduct comprehensive evaluations to demonstrate the robustness of FaceAdv.**

***Index Terms*— Adversarial examples, face recognition systems, adversarial stickers.**

I. Introduction

FACE recognition (FR) systems based on state-of-the-art deep neural networks (DNNs) have been widely used as a prominent biometric technique for authentication and identification in many application scenarios, such as payment authorization [36] and entry/exit management [34]. Compared with other approaches of recognition (e.g., passwords, smart cards, voiceprint, and fingerprints), facial characteristics of an individual are relatively more difficult to be stolen, forgotten, or replicated [41], and the recognition process can be carried out without any physical contact.

DNN-based FR systems, however, have been proved to be vulnerable to adversarial examples [6], which are derived from composing original images with *perturbations*, resulting in an incorrect output of FR systems. Adversarial attacks can be divided into two categories: *digital-world attacks*, where the attacker can feed manipulated digital images directly into DNNs, and *physical-world attacks*, where DNNs only accept inputs from a camera and the attacker can only present adversarial images to the camera. Despite the fact that digital-world attacks can achieve high performance, they cannot be successfully transferred to the physical world because of the dynamic physical conditions (e.g., different viewing angles and distances) [28].

The desired properties of physical-world attacks are *effectiveness*, *robustness* and *easiness*. Physical attacks aim at successfully cheating target FR systems and resisting to the variation of environmental conditions. The easiness means that the attacks should be easily launched without using customized equipment to generate perturbations in the physical world. Recent studies are devoted to designing practical attacks in the physical world [11], [26], [28], [43]. Based on the difference in image-forming principles between cameras and human eyes, some approaches [28], [43] can be used to avoid perturbations being observed. Perturbations are projected by specialized devices on faces (e.g., a cap with LEDs [43] or a projector [28]), which require considerable resources. Perturbations projected by LEDs can be easily filtered with infrared cut-off lens, making these attacks lose effectiveness. There are several approaches that paste adversarial stickers to the eyeglasses [26] or the cheek [11] and achieve a higher success rate. However, the generated adversarial examples are not robust enough, e.g.

Manuscript received March 10, 2021; revised May 21, 2021 and June 20, 2021; accepted July 10, 2021. Date of publication August 3, 2021; date of current version August 19, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1006101, in part by Beijing Nova Program under Grant Z201100006820006, in part by NSFC Projects under Grant 61972039 and Grant 61932016, in part by Beijing Natural Science Foundation under Grant 4192050, in part by the Open Research Projects of Zhejiang Laboratory under Grant 2020AA3AB04, in part by China National Funds for Distinguished Young Scientists under Grant 61825204, in part by Beijing Outstanding Young Scientist Program under Grant BJJWZYJH01201910003011, in part by ARC Discovery Grant under Project DP190103660 and Project DP200103207, and in part by ARC Linkage Grant under Project LP180100663. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andrew Beng Jin Teoh. *(Corresponding authors: Meng Shen; Liehuang Zhu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Beijing Institute of Technology.

Meng Shen is with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China, and also with Peng Cheng Laboratory (PCL), Shenzhen 518066, China (e-mail: shenmeng@bit.edu.cn).

Hao Yu is with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: csyuhao@bit.edu.cn).

Liehuang Zhu is with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: liehuangz@bit.edu.cn).

Ke Xu is with the Department of Computer Science, Tsinghua University, Beijing 100084, China, also with Beijing National Research Center for Information Science and Technology (BNRist), Beijing 100084, China, and also with Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: xuke@mail.tsinghua.edu.cn).

Qi Li is with the Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China (e-mail: qi.li@sz.tsinghua.edu.cn).

Jiankun Hu is with the School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2610, Australia (e-mail: j.hu@adfa.edu.au).

Digital Object Identifier 10.1109/TIFS.2021.3102492

their performance degrades significantly when the attackers do not look directly at the camera.

Inspired by the attack utilizing adversarial stickers to target at DNN-based recognition systems (e.g. traffic signs [6]) in physical scenarios, we utilize stickers pasted on human faces to attack FR systems. The sticker-based attacks are inherently easy to launch, as they do not require any specialized devices. We believe that although stickers are noticeable to human eyes, the sticker-based attacks can still take effect in unattended scenarios, such as unlocking a mobile phone or a car [33].

It is challenging to design effective and robust sticker-based attacks in the physical world. First, as an attacker needs to paste one or multiple stickers on the faces to perform attacks, it is of great importance to determine the critical regions where these stickers can be attached, which is referred to as *sticker localization* [13]. Adversarial examples placed only on the nose fail to attack FR systems [22]. To make the digitally designed stickers maintain effectiveness in the physical world, the perturbed face images taken by cameras in FR systems should be efficiently and accurately simulated. Second, in the physical world, it is usually difficult to paste stickers exactly on the same position as designed, and environmental conditions (e.g., user-camera distance, brightness, and head poses) always change, resulting in a severe impact on the performance of attacks, which is referred to as *perturbation loss* [28], [38].

In this paper, we propose FaceAdv, a novel method to realize effective, robust and easy physical-world attacks. FaceAdv designs a specific generator to craft adversarial stickers that are attached to regions chosen by the sticker localization. For the easiness of FaceAdv, the sticker localization algorithm limits the number and shape of stickers to reduce the preparation time and avoid other facial organs (e.g. eyes) being covered by the sticker corners.

For tackling the first challenge, we analyze the importance of different regions of human faces and choose five candidate positions (i.e., two superciliary arches, the nasal bone and two nasolabial sulcus) to attach adversarial stickers. Hence, FaceAdv will generate several stickers at a time and place them on the chosen regions of human faces, which can reduce the area of each crafted sticker while keeping the effectiveness. When training the generator, the face images with these stickers will be fed into the target FR system to check whether they successfully cheat or not. Inspired by the R-Net [5] to build accurate 3D face shape from a single image, we propose a new convertor in FaceAdv to digitally paste crafted adversarial stickers on human faces.

To deal with the second challenge, FaceAdv adopts three measures: 1) drawing samples (i.e. face images) from a distribution that models physical dynamics (e.g., varying distance, angles and ambient brightness), 2) rotating, scaling and translating stickers in the convertor to simulate errors when pasting them on real faces in the physical world, and 3) smoothing stickers with the total variation loss to minimize differences between adjacent pixels in stickers.

Extensive experiments are conducted to evaluate the performance of FaceAdv. The well-known face dataset LFW [10] along with a participant dataset VolFace are utilized to investigate the success rate of FaceAdv and a state-of-the-art method AGNs [27] on three typical FR systems (i.e., ArcFace [4], CosFace [35], and FaceNet [23]). The results show that both methods could achieve high success rate in digital scenarios. In physical scenarios, FaceAdv can significantly improve the success rate by a round 50% over AGNs, for both the dodging attacks and impersonating attacks. The robustness of FaceAdv is also confirmed as its success rate is kept at a high level when environmental conditions change.

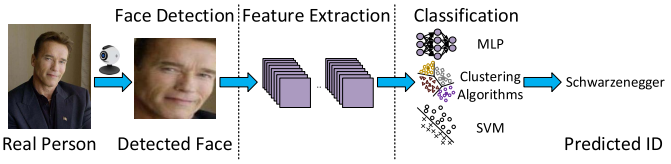


Fig. 1. The typical workflow of FR systems.

We summarize the main contributions as follows:

- We propose FaceAdv to craft several adversarial stickers with different shapes at a time. These stickers are flexibly attached to special positions (e.g., the nasal bone and the nasolabial sulcus) of human faces.
- We design a convertor to generate facial images with adversarial stickers while training the generator for simulating the capture of human faces with stickers attached in the physical world.
- We analyze three state-of-the-art FR systems (i.e. ArcFace, CosFace, and FaceNet) to find critical regions of human faces and conduct extensive experiments to demonstrate the effectiveness, robustness and transferability of the stickers in both the digital and physical scenarios.

This study was approved by the Institutional Review Board (IRB) of our university, and all the participants were provided written informed consent. We summarize the typical workflow of FR systems and existing attack algorithms in Section II before describing the threat model in Section III. After that, we introduce the overview of FaceAdv in Section IV and describe the design details in Section V. Next, we evaluate its effectiveness and robustness in Section VI. Finally, we make brief discussions in Section VII and conclude this paper in Section VIII.

II. Background and Related Work

In this section, we first introduce the typical workflow of FR systems based on DNNs and describe the state-of-the-art DNNs. Then, we summarize the recent achievements in the digital- and physical-world attacks on FR systems.

A. Background of FR Systems

We introduce the typical workflow of FR systems as illustrated in Fig. 1. It generally consists of three steps: face detection, feature extraction, and classification [7].

1) Face Detection: The camera embedded in an FR system takes images of one's face, which are then used to localize the facial region. The purpose of this step is to determine

TABLE I
SUMMARY OF TYPICAL ATTACKS ON FR SYSTEMS (FD FOR FEATURE DETECTION, FE FOR FEATURE EXTRACTION AND CF FOR CLASSIFICATION)

Domain	Attacks	Targeted Stages	Method Descriptions	Limitations
Digital	Bose et al. [1]	FD	The architecture of GANs to craft digital adversarial examples	Only working in the digital world
	Yang et al. [39]	FD	Automatically designing universal patches	
	A^3GN [37]	FE	A discriminator judging between original and adversarial images	
	Garofalo et al. [7]	CF	Deploying a poisoning attack to cheat SVM	
	Dabouei et al. [3]	FE & CF	Manipulating the locations of landmarks	
Physical	Kaziakhmedov et al. [11]	FD	Different face attributes printed by a white and black printer	Only attacking MTCNN
	IMA [43]	FE	Projecting the perturbations on human faces using infrared	Easily losing effectiveness with low-cost lens
	Komkov et al. [13]	FE	An adversarial sticker attached on the hat	Poorly performing in impersonating attacks
	Pautov et al. [22]	FE	An adversarial patch pasted on different areas	Only working when attackers looking directly
	AGNs [27]	FE & CF	Generating adversarial stickers attached on the eyeglasses	Easily being subject to the head pose
	VLA [28]	FE & CF	Hiding perturbations using PoV	Inconveniently launching attacks
	FaceAdv (this paper)	FE & CF	Crafting stickers with different shapes attached on critical regions	Only aiming at unattended scenarios

whether human faces exist in the image or not. When faces are detected, systems pre-process each face in the image to create the normalized and fixed-size input to the following DNNs for feature extraction. Since MTCNN [40] produces real-time and high-precision face detection results, we use it as the face detector in this paper.

2) Feature Extraction: It plays an important role in FR systems, which uses CNNs as the face embedding model to craft embedding vectors [29], [32]. The input is the fixed-size, frontalized face image and the output is the feature vector that describes the prominent features of the face image. Several state-of-the-art CNNs have been proposed for feature extraction:

- FaceNet [23], which directly learns a mapping from face images to a compact Euclidean space where the distance is a measure of face similarity.
- CosFace [35], which uses large margin cosine loss to maximize inter-class variance (i.e. the cosine distance variance of different persons) and minimize intra-class variance (i.e. the cosine distance variance of embedding vectors of the same person) in the angular space.
- ArcFace [4], which borrows the idea of CosFace and introduces an additive angular margin loss to obtain highly discriminative features for FR systems.

3) Classification: A low-dimensional representation created by the feature extractor can be efficiently used for classification. Based on application scenarios of FR systems, this step can be divided into the binary classification (i.e., authentication) and the multi-class classification [14]. Several classifiers are applied for either the binary classification, such as SVM [7], or the multi-class classification, such as Multi-Layer Perceptrons (MLPs) and K-Means.

B. Summary of Attacks on FR Systems

With the wide adoption of FR systems in various scenarios, their security and safety have attracted increasing research interests. Existing attacks on FR systems can be roughly classified into two categories: the *digital* attacks, which generate imperceptible perturbations added on digital face images, and the *physical* attacks, which design effective perturbations to mislead FR systems in the physical world.

Existing typical attacks on FR systems are summarized in Table I, which target at the different stages of an FR system. An adversary targeting at the face detector aims to mislead the face detector to avoid face locating. Attacks on feature extractors can reduce the distance of low-dimensional vectors of two images from different persons or enlarge the distance of two images from the same person. The rest of the attacks can also cheat classifiers to make incorrect decisions.

1) Digital Attacks: These methods directly manipulate the pixel values of face images and feed the modified images into FR systems. Besides the effectiveness, the imperceptibility of pixel-level perturbations is also critical. Thus, these algorithms always constrain the magnitude of perturbations to ensure the perturbed images visually similar to the original images.

As illustrated in Table I, there are some approaches to craft digital adversarial examples for the face detector [1], [39], the feature extractor [3], [37] and the classifier [3], [7].

In most cases, however, an adversary cannot directly manipulate the input images to FR systems [30], making these attacks unapplicable in the physical world.

2) Physical Attacks: Due to spatial constraints (e.g., varying ambient brightness and face posture), fabrication errors and resolution changes, perturbations working well in the digital world will lose effectiveness. Recent studies focus on designing physical attacks to generate robust perturbations that can survive in the physical world.

In order to achieve effectiveness and imperceptibility simultaneously, several attacks utilize the difference in image-forming principles between cameras and human eyes. Zhou *et al.* [43] deceived the feature extractor by illuminating the subject using infrared, as infrared-based perturbations are invisible to human eyes but can be captured by cameras. However, these perturbations are easily filtered out by infrared cut-off filters that are commonly equipped in solid state cameras (e.g., CMOS). Shen *et al.* [28] proposed VLA, which leverages visible light to generate a perturbation frame and a concealing frame that are alternately projected on human faces. Its imperceptibility relies on a phenomenon called Persistence of Vision: if the two frames change faster than 25Hz, human brain will mix them together and thus cannot observe perturbations. However, it requires certain kinds of equipment and cannot be easily conducted in real-world scenarios.

Several studies pay much attention to improving the convenience and effectiveness of adversarial perturbations rather than keeping their imperceptibility. Kaziakhmedov *et al.* [11]

proposed different face attributes printed by an ordinary white and black printer and attached them to either the medical face mask or the real face to attack the MTCNN face detector. Komkov *et al.* [13] elaborately designed an adversarial sticker attached on the hat to cheat ArcFace the feature extractor. Unlike the two methods above, Sharif *et al.* [26], [27] presented white-box attacks to generate adversarial stickers attached on the eyeglasses to cheat the feature extractor and the MLP classifier. They tried their best to make these adversarial stickers inconspicuous, other than imperceptible, to human eyes. However, colorful frames of eyeglasses still look unusual because frames are generally in solid color in daily life.

FaceAdv proposed in this paper aims at generating effective and robust adversarial stickers to cheat the feature extractors and the MLP classifiers. To mitigate the perturbation loss, FaceAdv applies several stickers (i.e., three in our implementation) attached on critical regions of human faces to attack FR systems in dodging and impersonating attacks. FaceAdv introduces a series of transformations to simulate the digital-to-physical transformation process so as to resist the variation of environmental conditions (especially the head pose). To make the stickers relatively normal, FaceAdv provides the stickers with different shapes and the area of the stickers is smaller than that of the ones crafted by Komkov *et al.* [13].

III. Threat Model

In this section, we describe the threat model and design goals of FaceAdv.

A. White/Black-Box Assumption

There are two typical scenarios for adversarial attacks: white-box scenarios and black-box scenarios.

In white-box scenarios [26], [27], the adversary has full knowledge of the target FR system, including the dimension of input, the architecture and parameters of the feature extractor and the MLP classifier. The adversary can analyze the vulnerability of the operating system that FR systems are deployed in. Further, the adversary hijacks these systems to obtain desired models.

Black-box scenarios assume that the adversary has no access to the target FR system. He can train a local substitute model as the attacked model [20]. Then, adversarial examples can be generated by attacking the substitute model to deceive the target FR system.

We also assume the FR system accessed by the adversary is already well trained so that the adversary cannot manipulate the training process of the system. Thus, the poisoning attack, which requires injecting adversarial images in the training set of FR systems, is beyond our consideration in this paper.

We select three state-of-the-art FR systems with different feature extractors as the target models, i.e., FaceNet [23], CosFace [35], and ArcFace [4], as described in Section II-A.

B. Attack Goals

The goal of the adversary is to trick FR systems to misclassify the adversarial input. Given an FR system $\mathcal{F}_\theta(\cdot)$ with parameters θ containing the feature extractor and the classifier,

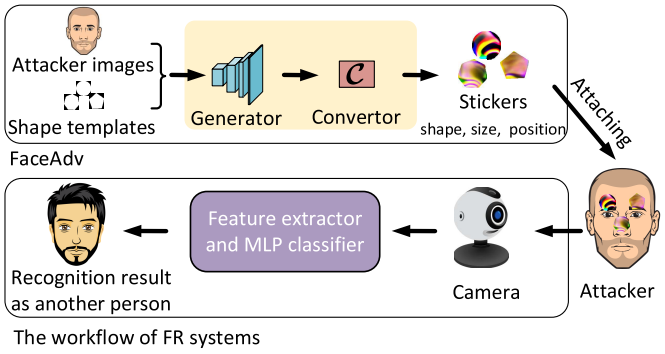


Fig. 2. The physical-world attacks on FR systems using adversarial stickers crafted by FaceAdv.

and an input facial image x with its ground truth label y (e.g., identity), an ideal FR system can label x as y, which is defined as $\mathcal{F}_\theta(x) = y$. However, the adversarial image can cause the system to make an incorrect prediction. In this paper, we consider two types of attacks.

1) Dodging Attacks: The adversary aims at crafting the adversarial image $x^* = x + \Delta_x$ with the perturbation Δ_x to mislead the classification result, which can be expressed by $\mathcal{F}_\theta(x^*) \neq y$. For instance, the adversary can be a terrorist who needs to bypass an FR system for biometric security checking.

2) Impersonating Attacks: The adversary attempts to mislead the FR system by classifying the perturbed image $x^* = x + \Delta_x$ as a target label y^*, i.e., $\mathcal{F}_\theta(x^*) = y^*$. In real-world applications, y^* can be a legitimate individual with a certain authority. Such an attack enables the adversary to illegally unblock authentication.

In this paper, our primary goal is to generate adversarial examples on the *effectiveness* to deceive the state-of-the-art FR systems rather than on the inconspicuousness of adversarial examples. The latter might not be necessary in certain unattended scenarios, such as unlocking a mobile phone or a car [33], the face scan payment in unattended convenience stores [17], and the access control of smart buildings [19].

IV. Overview of FaceAdv

In this section, we first describe the overview of FaceAdv followed by a case study to show the adversarial examples against the target FR systems.

When designing FaceAdv, we desire that it will be launched easily in the physical world (i.e., easiness). Therefore, we resort to adversarial stickers that can be simply stuck onto human faces, which requires no special devices such as the LEDs or projectors [28], [43]. Considering the perturbation loss from the digital to the physical world, it is a challenging task to make the sticker-based adversarial examples effective and robust in varying environmental conditions.

The workflow of FaceAdv is illustrated in Fig. 2. The adversary takes several photos of his own face in different environmental conditions (e.g., user-camera distance, brightness and head pose) and chooses the desired locations and shapes of the stickers to train the generator. During this training, the convertor generates facial images with stickers

TABLE II
SAMPLES OF DODGING AND IMPERSONATING ATTACKS

Attacker	Mode	Target	Target Model		
			ArcFace	CosFace	FaceNet
	Dodging	Another person			
	Impersonating				

to simulate the capture of human faces with stickers attached in the physical world. Then, the adversary utilizes the trained generator to craft adversarial stickers (including shapes, sizes and positions) and prints these stickers on paper. Finally, the adversary attaches the printed stickers to his own face and launches an attack in front of the camera.

To achieve higher effectiveness, FaceAdv can craft several stickers pasted on different regions of a human face. In general, the more stickers pasted on a face, the higher success rate FaceAdv has. In an extreme case, the entire face is covered by crafted stickers, which looks like that the attacker wears a mask. However, the face liveness detection in FR systems can discover mask-based attacks [12]. Thus, we should limit the area of stickers to pass the face liveness detection and to increase the inconspicuousness of the attack.

Next, we give an example to launch dodging and impersonating attacks on target FR systems using adversarial stickers crafted by FaceAdv, as illustrated in Table II. In dodging attacks, the victim is another person other than the attacker. However, in impersonating attacks, stickers can ensure the recognition result is the targeted victim (e.g., Keanu Reeves in the 3rd column of Table II) other than the attacker.

V. DETAILS OF FACEADV

In this section, we introduce design details of FaceAdv, including training the perturbation generator and digitally attaching stickers to human faces.

A. FaceAdv Architecture

As mentioned above, stickers crafted by FaceAdv should be pasted on human faces, which raises three challenging problems. 1) Since traditional stickers are usually in the form of rectangles and squares [13], [22], the four corners of a sticker may cover facial organs when it is attached to specific locations (e.g., the nasal bone). 2) It is crucial to determine the positions where these stickers can be pasted. Pautov *et al.* [22] place crafted adversarial stickers onto difference positions of the human face (i.e., eyeglasses, forehead and nose) to attack the FR system and find the positions of stickers have a dramatic influence on the effectiveness. 3) How to efficiently and accurately obtain the camera-perceived face images when stickers are pasted on real human faces.

In this paper, we design a FaceAdv architecture to craft adversarial stickers, which is composed of a sticker generator $\mathcal{G}$ and a convertor $\mathcal{C}$ to separately tackle the first challenge and the other two challenges mentioned above. The aim of the generator is to craft stickers, while the convertor is utilized to digitally paste stickers to human faces for imitation. The face image with stickers is fed into the target FR system to find the optimization direction during the training phase.

For the sake of easiness, FaceAdv limits the number and shape of stickers to shorten the preparation time for an attack and to avoid other facial organs (e.g., eyes) being covered by the four corners of stickers. The sticker generator can generate adversarial stickers and shapes with four corners cut off by GANs. GANs, which casts generative modeling as the two-player game between a pair of generator $\mathcal{G}$ and discriminator $\mathcal{D}$, are a powerful class of generative models. It generally takes a lot of time to train the generator in the two-player game. To address this issue, we divide the generator into a *sticker* component and a *shape* component, where the former is applied to craft the content of stickers while the latter generates the shape of stickers. Further, we can use the shape datasets to pre-train the sticker component and load the pre-trained parameters to reduce the training time. The two components share the parameters of the two previous layers, which further lowers the computational complexity on the side of the user.

FaceAdv utilizes crafted shapes to tailor the stickers fabricated by the generator and the convertor to digitally attach them to the face images that are fed into the FR system. Then the generator updates the parameters according to the recognition results. In the next subsections, we will describe the sticker generator and the convertor.

B. Sticker Generator

Inspired by producing adversarial stickers placed on eyeglasses to cheat FR systems in the real world [27], we introduce GANs to generate stickers. The difference lies in that we resort to GANs to constrain the shape, rather than the content, of adversarial stickers. Since adversarial stickers crafted by previous approaches are always square or rectangular [13], [22], different shapes of stickers will probably reduce the total size that, in turn, is harder to detect. In this work, we employ GANs to craft adversarial stickers with different shapes in Fig. 3. The notations used in the rest of this paper are summarized in Table III.

Unlike traditional GANs, there are three networks in FaceAdv. The generator consists of three branches, each of which is utilized to generate a shape mask and an adversarial sticker. The shape mask is a binary image that only has two colors (black and white), which is used to tailor the original square stickers so that the shape of the cropped stickers is the same as that in the training datasets (specifically, the shape template). The goal of the discriminator is to distinguish the shapes crafted by the generator for constraining that the crafted shapes are all in the training datasets. There is another network (i.e., the target FR system $\mathcal{F}$) to obtain the recognition results, whose parameters are frozen in the process of training the generator and the discriminator. The aim of the target FR system is to inform the generator of the effectiveness of the crafted stickers so that it can adjust the optimization direction in time during the training process.

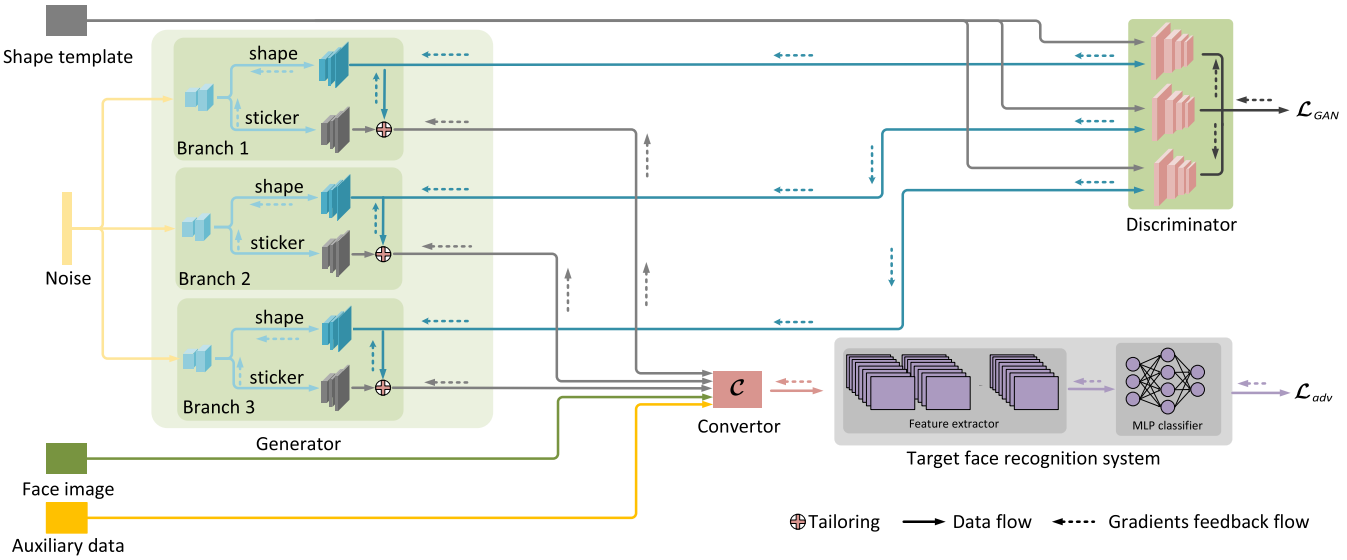


Fig. 3. The FaceAdv architecture. The generator is divided into three branches for crafting three stickers, and the discriminator is utilized to judge whether the input shape is created by the generator or not. The stickers are fabricated by tailoring the original square rectangular according to the crafted shapes. Then, the tailored stickers are digitally applied to human faces using the original images and auxiliary information, which are finally fed into the FR system to obtain the recognition result.

TABLE III
NOTATIONS USED IN THIS PAPER

Notation	Description
$\boldsymbol{x}_A$	A face image of user A
P_A	The label of user A
γ_i	The parameters of the illumination model, $i \in [1, 27]$
$\boldsymbol{n}$	Noise sampled from the normal distribution
$\mathcal{G}$	The generator to craft stickers with different shapes
$\mathcal{D}$	The discriminator to discriminates between crafted shapes and shape templates
$\mathcal{F}$	The target face recognition system
$\mathcal{C}$	The convertor to digitally attach stickers to faces
$\mathcal{L}_{GAN}$	The adversarial loss of GANs
$\mathcal{L}_{adv}$	The loss for fooling the face recognition system

FaceAdv can generate multiple stickers at a time. The number of stickers is the same as that of the branches in Fig. 3. Based on our preliminary experiments, we set the number of stickers as 3 (i.e., Branch 1 to 3) by default, because it can achieve a better balance between effectiveness and easiness. It is flexible to adjust the number of adversarial stickers in FaceAdv: we can add (or remove) branches to (or from) the generator and the discriminator and then alter the sticker positions of the convertor.

Since FaceAdv tailors some parts of square stickers to form stickers with different shapes, it is important to design a reasonable mechanism to assure the cut part cannot largely damage the effectiveness of stickers. In Fig. 3, the gradients from the discriminator can place restrictions on the shapes of stickers. Besides, the gradients from the target FR system are applied to change the parameters of the generator containing parts for crafting shapes and stickers. Apparently, the gradients flowing into the part of creating stickers can alter the content of stickers, and the gradients flowing into that of producing shapes can also change the shapes of stickers. In this way, if FaceAdv tailors some important regions, L_{adv} that indicates the effectiveness of adversarial stickers will be larger and the generator will realize this problem and update the parameters.

To improve the robustness of crafted stickers, when training the generator, the face images are composed of multiple images captured in different conditions so that stickers can work in physical scenarios.

There Are Two Loss Functions: $\mathcal{L}_{GAN}$ and $\mathcal{L}_{adv}$, where $\mathcal{L}_{GAN}$ is utilized to train the shape branch of the generator and the discriminator, and $\mathcal{L}_{adv}$ is applied to optimize the sticker branch of the generator. They will be further discussed in Section V-D.

There is another key part of FaceAdv to be elaborated: the convertor, which connects the generator and the target FR system. As mentioned previously, FaceAdv applies 3D face shapes to digitally paste adversarial stickers to human faces and the process of attaching stickers should be differentiable so that the gradients from the target FR system can smoothly flow into the generator. The auxiliary data is used in the convertor, which includes the 3D face shape and the parameters of illumination models as described in the next subsection.

C. Digital Transformation of Physical Sticker

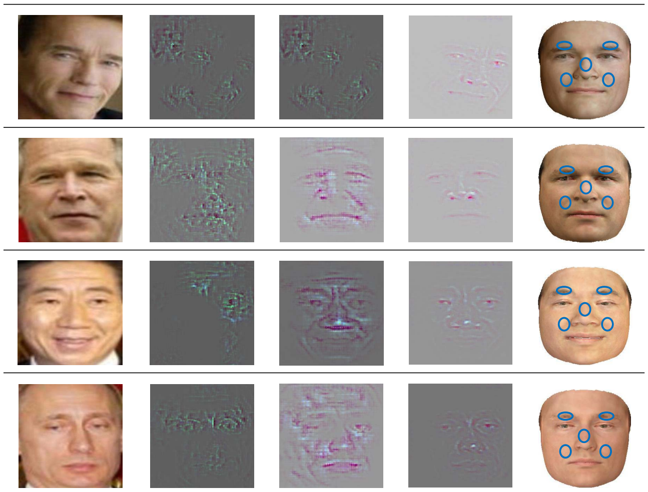
As described above, the convertor digitally attaches adversarial stickers to the critical regions of face images. In this subsection, we will elaborate on the design of the convertor.

1) Sticker Localization: The locations of stickers on real faces can largely affect the attack effectiveness. A natural idea is to place the stickers on the positions where the target FR system extracts discriminative features of human faces.

To investigate regions of human faces where each target FR system extracts features, we leverage Guided Grad-CAM [24]

TABLE IV
THE LOCALIZATION MAPS CRAFTED BY GUIDED GRAD-CAM

Original Images	Model			3D Faces
	ArcFace	CosFace	FaceNet	



to analyze effective regions of the three FR systems, as illustrated in Table IV. The Guided Grad-CAM uses the gradient information flowing into the last convolutional layer of the feature extractor to analyze the importance of each region in the image for a decision of interest.

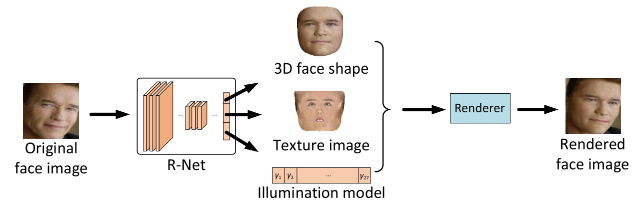
The localization maps of four individuals[1] in Table IV show that these FR systems invariably focus on regions near to the facial organs (e.g., eyes, nose, and mouth) to make a decision. In ArcFace, the regions to extract features are not always near to facial organs (e.g., temples of Schwarzenegger in Table IV). CosFace and FaceNet extract valuable features near to the nasolabial sulcus.

Due to the face liveness detection and the convenience of tailoring adversarial stickers, it is important to balance the number of stickers and the area of each sticker. When fixing the total area of perturbed regions on human faces, if the number of stickers is larger, the area of each sticker will be smaller, but the time to tailor stickers will be longer. We choose three adversarial stickers to achieve a better balance.

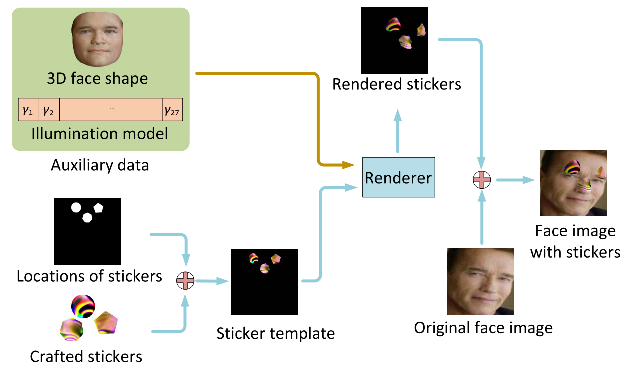
The stickers crafted by FaceAdv cannot directly cover facial organs, so FaceAdv tends to attach the stickers in regions near to facial organs. Based on the observations above, we select five regions (i.e., two superciliary arches, two nasolabial sulcus, and the nasal bone) as the candidate regions for pasting adversarial stickers generated by FaceAdv, as marked by the blue circles in the last column of Table IV. Existing studies have also shown the effectiveness of adversarial stickers pasted on these regions, e.g., two superciliary arches are selected by [13], [27] and the nasal bone is chosen by [22]. We will describe the selection of the final regions to attach stickers for each target FR system in Section VI-B.

2) Attaching Stickers: When training the generator, parameters will be updated according to the recognition results of human faces with created stickers. Hence, the perturbed

[1]They are Arnold Schwarzenegger, George W Bush, Roh Moo-hyun and Vladimir Putin.



(a) The workflow of rendering the texture images



(b) The proposed RSO for digitally attaching stickers to faces

Fig. 4. The implementation of the convertor $\mathcal{C}$. It utilizes R-Net to generate auxiliary information and applies the workflow of rendering texture images to digitally paste adversarial stickers to human faces.

face images taken by cameras in the real world should be effectively and accurately simulated.

Recent studies, which exploit stickers to deceive FR systems, proposed algorithms to digitally paste these stickers to human faces or facial accessories, which simulates the appearance of human faces with stickers in the real world [13], [22], [27]. However, these algorithms have apparent limitations: they can neither handle the situation where the attacker does not face directly [22], [27], nor digitally attach different shapes of stickers to human faces [13]. The stickers crafted by FaceAdv have various shapes so that we have to design a new method to attach stickers to faces digitally.

We employ the 3D face reconstruction method [5] to estimate 3D face shapes, illumination model as well as the camera model, as illustrated in Fig. 4(a), and leverage the resulting face shapes to digitally attach adversarial stickers to human faces. After getting these information, the differentiable renderer can render the texture image according to the 3D face shape while fusing the reckoned ambient brightness.

Motivated by the remarkable performance of R-Net [5], we propose a new method named *rendering stickers only* (RSO) to render the texture image with only stickers. The proposed RSO method in Fig. 4(b) replaces the texture image with the location image of stickers that indicates locations to place these stickers. The conversion from texture image to location image follows this rule: the pixel value in the regions of the original texture image is set to 0 while that in the regions for pasting stickers is equal to 255. After putting these stickers on the chosen positions, the renderer can craft an image containing the stickers only, and the resulting image can cover the original face image to produce the face image with stickers.

The rendered face image created by RSO retains many realistic details and thereby is more similar to the original face image. FaceAdv adopts RSO to digitally attach adversarial stickers to human faces in the convertor. To reduce time consumption, FaceAdv generates the 3D face shape, the location image of stickers and the parameters of the illumination model using the R-Net in advance. The convertor gains the auxiliary information from the input when training the generator.

D. Loss Functions

Having the architecture of FaceAdv, in this subsection, we will design loss functions for the two attack modes (i.e. dodging attacks and impersonating attacks) and formally describe the training algorithm.

Although the GAN has achieved appealing results in the image generation, training GANs stably is still a challenging problem. To alleviate this problem, we resort to the Wasserstein GAN with gradient penalty (WGAN-GP) [8]. Essentially, the goal of GANs is to transform the distribution of a random noise $\boldsymbol{n}$ into the distribution of the input data (i.e. the shape image in this paper). WGAN-GP utilizes the discriminator to calculate the Wasserstein distance between the shapes crafted by the generator and the shape templates, which is denoted by $\mathcal{L}_{GAN}$ in Fig. 3. $\mathcal{L}_{GAN}$ can be logically decomposed as $\mathcal{L}_{\mathcal{D}}(\boldsymbol{n}, \boldsymbol{s})$ and $\mathcal{L}^{s}_{\mathcal{G}}(\boldsymbol{n})$, which can be used to train the generator for crafting different shapes.

The discriminator aims to distinguish between crafted shapes and shape templates. When feeding a shape into the discriminator, it will output a number: the smaller number is more likely to be judged as a crafted shape; otherwise, it is recognized as a shape template. Thus, the objective of the discriminator is minimizing $\mathcal{L}_{\mathcal{D}}$.

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\boldsymbol{n}, \boldsymbol{s}) = &\mathbb{E}_{\boldsymbol{n}\sim\mathbb{P}_{\boldsymbol{n}}}(\mathcal{D}(\mathcal{G}(\boldsymbol{n}))) \\ &-\mathbb{E}_{\boldsymbol{s}\sim\mathbb{P}_{\boldsymbol{s}}}(\mathcal{D}(\boldsymbol{s})) \\ &+\lambda\mathbb{E}_{\hat{\boldsymbol{s}}\sim\mathbb{P}_{\hat{\boldsymbol{s}}}}[(\| \bigtriangledown_{\hat{\boldsymbol{s}}} \mathcal{D}(\hat{\boldsymbol{s}})\|_2 - 1)^2] \end{aligned} \tag{1}$$

where $\mathbb{P}_{\boldsymbol{n}}$ and $\mathbb{P}_{\boldsymbol{s}}$ are the distribution of the noise $\boldsymbol{n}$ (i.e. the normal distribution) and the distribution of the shape template $\boldsymbol{s}$, respectively. The last term of Eq. (1) is the gradient penalty, which makes training more stable. $\mathbb{P}_{\hat{\boldsymbol{s}}}$ is sampling uniformly between pairs of images sampled from the shape template distribution $\mathbb{P}_{\boldsymbol{s}}$ and the crafted shape distribution $\mathbb{P}_{\mathcal{G}(\boldsymbol{n})}$, which can be formulated as:

$$\hat{\boldsymbol{s}} = \epsilon\boldsymbol{s} + (1-\epsilon)\mathcal{G}(\boldsymbol{n}) \tag{2}$$

where $\boldsymbol{s} \sim \mathbb{P}(\boldsymbol{s})$, $\mathcal{G}(\boldsymbol{n}) \sim \mathbb{P}_{\mathcal{G}(\boldsymbol{n})}$ and $\epsilon \sim U[0, 1]$.

The goal of the generator is to mislead the discriminator by labeling crafted shapes as shape templates, which can be expressed in Eq. (3).

$$\mathcal{L}^{s}_{\mathcal{G}}(\boldsymbol{n}) = -\mathbb{E}_{\boldsymbol{n}\sim\mathbb{P}_{\boldsymbol{n}}}(\mathcal{D}(\mathcal{G}(\boldsymbol{n}))) \tag{3}$$

Next, we focus on the loss function $\mathcal{L}_{adv}$, which represents either the opposite of the distance between predicted and ground-truth classes in dodging attacks, or the distance between predicted and target classes in impersonating attacks. In black-box scenarios, transferability is a desirable property for adversarial examples, which means that adversarial examples generated for a certain FR system can fool the other FR systems with different architectures. To improve the transferability, we use the ensemble of source FR systems to train the sticker generator [18].

In dodging attacks, the attacker P_A aims to deceive the FR system by misclassifying himself as another person P_B ($P_A \neq P_B$). Thus, FaceAdv should reduce the probability of class P_A and make it less than the probability of another class,

$$\mathcal{L}_{adv}(\boldsymbol{x}_A, \boldsymbol{n}, \mathcal{A}, P_A) = -\frac{1}{m}\sum_i \mathcal{F}_i(\mathcal{C}(\boldsymbol{x}_A, \mathcal{G}(\boldsymbol{n}), \mathcal{A}), P_A) \tag{4}$$

where $\mathcal{F}_i(\cdot, \cdot)$ represents the cross entropy that is commonly used in image classification [27], m is the number of elements in the ensemble of source FR systems and $\mathcal{A}$ is the auxiliary data for the convertor. In particular, m is set to 1 in white-box scenarios. By minimizing Eq. (4), the output of the cross-entropy function will increase so that the probability of the class P_A can be reduced.

In impersonating attacks, we expect that FR recognizes the identity of the attacker P_A as the target class P_B, which can be formulated by Eq. (5).

$$\mathcal{L}_{adv}(\boldsymbol{x}_A, \boldsymbol{n}, \mathcal{A}, P_B) = \frac{1}{m}\sum_i \mathcal{F}_i(\mathcal{C}(\boldsymbol{x}_A, \mathcal{G}(\boldsymbol{n}), \mathcal{A}), P_B) \tag{5}$$

The color change of neighboring positions in adversarial stickers will affect the perturbation loss. These images captured by cameras in the real world comprise smooth and consistent patches, where colors change gradually [26]. Due to this phenomenon, extreme difference between adjacent pixels in adversarial stickers cannot be accurately captured by cameras. Consequently, we use the total variation loss to smooth these stickers, which can be defined in Eq. (6),

$$\begin{aligned}\mathcal{L}_{tv}(\boldsymbol{n}) = \sum_{i,j}\big((\mathcal{G}(\boldsymbol{n})_{i,j} - \mathcal{G}(\boldsymbol{n})_{i,j+1})^2 \\ +(\mathcal{G}(\boldsymbol{n})_{i,j} - \mathcal{G}(\boldsymbol{n})_{i+1,j})^2\big)^{\frac{1}{2}}\end{aligned} \tag{6}$$

where $\mathcal{G}(\boldsymbol{n})_{i,j}$ is the pixel in $\mathcal{G}(\boldsymbol{n})$ at coordinates (i, j).

In practice, printers used for printing adversarial stickers may only contain a subset of the $[0, 1]^3$ RGB color space (i.e., the color gamut $P \subset [0, 1]^3$). Thus, we employ the non-printability score (NPS) [26] to constrain the color of stickers in the color gamut. The NPS $\mathcal{L}_{nps}$ is defined in Eq. (7),

$$\mathcal{L}_{nps}(\boldsymbol{n}) = \sum_{\hat{p}\in\mathcal{G}(\boldsymbol{n})}\left[\prod_{p\in P}|\hat{p}-p|\right] \tag{7}$$

where $\hat{p}$ is the pixel in $\mathcal{G}(\boldsymbol{n})$. If $\hat{p}$ belongs to P, or is quite close to a certain $p \in P$, $\mathcal{L}_{nps}$ will be minimized.

Since the generator aims to craft stickers with different shapes to cheat the FR system, the entire loss of the generator is defined in Eq. (8),

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}^{s}_{\mathcal{G}} + \alpha\mathcal{L}_{adv} + \beta\mathcal{L}_{tv} + \gamma\mathcal{L}_{nps} \tag{8}$$

where α, β and γ are weights that control the relative importance of each objective. When α is large, the shape of the crafted stickers will be uncontrollable and is difficult to be

cut out in the real world. A larger β or γ means the color in the generated stickers tends to be the same (i.e., printable).

VI. Performance Evaluation

In this section, we comprehensively evaluate the performance of the proposed FaceAdv against three state-of-the-art FR systems. We are dedicated to answering the following questions: 1) How does FaceAdv generate appropriate stickers for each target FR system? 2) Will FaceAdv outperform the other methods when launching dodging attacks and impersonating attacks? 3) How will potential influencing factors affect the performance of FaceAdv? 4) Are the adversarial stickers crafted by FaceAdv transferable in black-box scenarios?

A. Experimental Settings

1) Testbed: We select ArcFace [4], CosFace [35], FaceNet [23] and VGGFace [21] as the state-of-the-art feature extractors in FR systems. Given a specific feature extractor, the corresponding MLP classifier of the target FR system needs to be trained. We use a server with 32GB RAM, Nvidia RTX 2070 GPU and AMD Ryzen 7 7200X CPU for all the training tasks. The target FR systems are deployed on a PC with 16GB RAM and Intel Core i7-9750H CPU. The camera is Logitech C270 and captures images with 960×1280 in pixels. The printer for making physical stickers is HP DeskJet 2677 and the printable colors consist of the same 30 colors as Sharif *et al.* [26]. The light source is Philips 66135, which can change the ambient light to 30lux, 130lux and 250lux, accordingly.

The default values of the experimental parameters are set as follows: user-camera distance of 50cm, sticker size of 90×90 in pixels, ambient brightness of 130lux, $r = 5e^{-4}$, $\alpha = 100$, $\beta = 1$, $\gamma = 10$ and a head pose of facing directly forward. All the experiments are conducted using the default settings unless otherwise specified.

2) Baselines: The target stage of the two algorithms [13], [22] in Table I is only the feature extractor, while that of FaceAdv is the feature extractor and the classifier. Since the target stage of FaceAdv is the same with AGNs [27], we employ AGNs as the state-of-the-art physical-world attack for comparison. Both FaceAdv and AGNs use sticker-like perturbations and adopt the similar methodology (i.e., GANs) in generating perturbations. According to the original paper, the parameter κ is set to 0.25. AGNs is evaluated with VGGFace and achieves the same performance as its original paper [27] for both digital and physical attacks.

3) Datasets: The well-known face dataset LFW [10] is a public benchmark for FR systems, which consists of $13,233$ images from $5,749$ individuals. For ArcFace [4], CosFace [35], FaceNet [23] and VGGFace [21], their pre-trained models are publicly available, achieving an accuracy of 99.65%, 99.23%, 99.65%, 97.22% on LFW, respectively. As mentioned above, we need to train an MLP classifier, which follows a feature extractor to form a complete FR system. However, we find that 4,069 labels in LFW only have a single image and thus cannot be used to train the MLP classifier. Therefore, we only select the labels with more than 10 images, which results in a subset $\mathbf{LFW}^-$ containing 143 labels as the *victim* dataset.

TABLE V
Details of Datasets

Dataset	# Labels	# Images	ArcFace	CosFace	FaceNet	VGGFace
LFW	$5,749$	$13,233$	99.65%	99.23%	99.65%	97.22%
LFW^-	143	$4,174$	97.39%	97.26%	99.70%	98.54%
VolFace	20	900	99.50%	99.84%	100.00%	100.00%

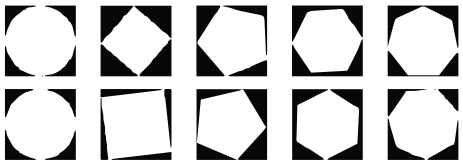


Fig. 5. Sampled shapes in VShape.

For launching physical-world attacks, there are 20 participants (including 12 males and 8 females) aged between 20 and 25 voluntarily participating in collecting the *attacker* dataset **VolFace**, which contains 45 face images for each participant. All the participants are Asian and wear no glasses or contact lenses. Our study is approved by the university IRB and we obtain written informed consent from all the participants. Then, we randomly select 80% of the images of each person in LFW^- and VolFace to train the MLP classifier, and the rest for testing the target FR systems. The accuracy of FR systems is shown in Table V.

We also have a *shape* dataset **VShape** to train the generator to fabricate different shapes, which contains five different shapes (i.e., circle, square, pentagon, hexagon and heptagon). This dataset consists of $15,000$ images and some samples in VShape are illustrated in Fig. 5. There is a trick that there should not be a big difference between the area of different shapes in the shape dataset.

4) Performance Metrics: In real-world applications, it is normal to record a video clip and extract several frames from this clip at random to FR systems [27]. Consequently, for each attack, we record a video of 25 seconds and randomly select 135 frames on average to obtain the classification result of the target FR systems. During face detection, only the face with the maximal confidence in one frame is applied to the recognition. For *dodging* attacks, the success rate is the fraction of faces that are classified as a different person from the attacker, and for *impersonating* attacks, it is defined as the fraction of faces that are classified as the target person.

B. Evaluation of Sticker Settings

The settings of adversarial stickers (e.g., locations and sizes) are critical to determine the performance of FaceAdv. Since launching physical-world attacks are more time-consuming, we resort to digital attacks in the following experiments: the adversarial images are created by attaching stickers on face images of the attackers in VolFace (i.e., 45 images for each attacker) and then used to cheat the target FR systems. Only the impersonate attack mode is involved as it is more challenging than the dodging attack mode.

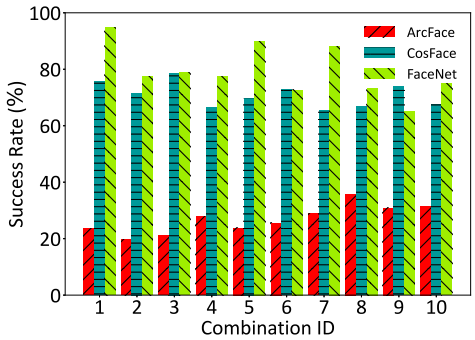


Fig. 6. Success rates of different location combinations.

TABLE VI
SUCCESS RATES (%) WITH VARYING STICKER SIZES

Target Model	Sticker Size (px)		
	80×80	90×90	100×100
ArcFace	16.89	35.78	56.67
CosFace	58.00	78.67	89.33
FaceNet	89.78	94.89	99.78

1) Locations of Stickers: As mentioned in Section V-C.1, we use Guided Grad-CAM to select five critical regions for attaching adversarial stickers. Since FaceAdv can generate three different stickers at a time, there are 10 combinations[2] of candidate positions to paste stickers. We randomly select five attackers from VolFace and two victims from LFW$^-$, and test each combination in turn for each attacker-victim pair. Thus, there are altogether 4,500 tests for each target FR system. In these experiments, the sticker size is fixed to 90px $\times$ 90px.

The success rates of FaceAdv are summarized in Fig. 6. In general, FaceNet is more vulnerable than the other two FR systems. We also observe that the combination that achieves the highest success rates varies a lot among different FR systems. Based on the results, we opt combination #8 (i.e., left superciliary arch, nasal bone, left nasolabial sulcus), #3 (i.e., right superciliary arch, left superciliary arch, left nasolabial sulcus), and #1 (i.e., right superciliary arch, left superciliary arch, nasal bone) as the best choices for attacking ArcFace, CosFace and FaceNet, respectively.

This result can be explained by the localization maps in Table IV. FaceNet always pays more attention to facial organs (e.g., the eyes, the nose and the mouth) and ignores the nasolabial sulcus, which is exactly the positions of combination #1. In contrast, both ArcFace and CosFace obtain information from the nasolabial sulcus. ArcFace tends to extract face features from partial face regions, which reduces the effectiveness of stickers. In the rest of the experiments, we regard these position combinations as the default settings.

2) Size of Stickers: We investigate how the size of stickers affects the performance of FaceAdv. Since the face size of each person is not necessarily the same and faces of different persons will be scaled to the same resolution in the sticker template, we measure the sticker size by pixel (i.e., px). The size of stickers in the final face images will vary with the head pose, as illustrated in Fig. 4(b), and thus sticker size refers to the size in the sticker template. We use the same attacker-victim pairs as those in the locations of stickers in Fig. 6, and consider three sizes: 80px $\times$ 80px, 90px $\times$ 90px and 100px $\times$ 100px. For each target FR system, we totally conduct 1,350 tests.

In our implementation, the size of sticker templates is 600px $\times$ 600px. Therefore, the total area of adversarial stickers accounts for less than 10% that of sticker templates.

The results are shown in Table VI. In general, a larger sticker helps to improve the success rate. In the physical world, however, the length of nasal bone and the distance between facial features limit the size of stickers. If the size of stickers pasted on the nasal bone is too large, it will cover eyes and cannot commendably handle the arc between the nasal bone and the glabella, which will significantly degrade the easiness of FaceAdv. Due to this limitation, we select 90px $\times$ 90px as the default sticker size.

3) Time Efficiency: In order to reduce the training time of the generator, we apply VShape to pre-train the generator and the discriminator only through $\mathcal{L}_{GAN}$ so that the shape component of the generator can generate different shapes before the generator is trained to attack FR systems.

With the fixed location and size of stickers, FaceAdv can train the generator only once to generate adversarial stickers for a specific FR system. In our tests, FaceAdv takes less than 26 minutes on average to train the generator and less than two seconds to craft adversarial stickers. When upgrading GPU to Nvidia RTX 2080Ti, the corresponding time is reduced to less than 20 minutes and less than one second, respectively.

C. Evaluation of Dodging and Impersonating Attacks

In Table II, we give several adversarial examples for dodging and impersonating attacks against the target FR systems. In this subsection, we will conduct extensive experiments to evaluate the effectiveness of FaceAdv.

The method AGNs utilizes 7 green marks to indicate the location of eyeglasses, and it cannot work well when an attacker does not look straight ahead, because the green marks may be cut off by the face detector. To make a fair comparison, we take another 45 face images for each participant who looks straight ahead, and these images are only used to calculate the success rate in the digital world. In the physical world, the user-camera distance is 50cm, the ambient brightness is 130lux, and the head pose is straight ahead.

1) Dodging Attacks: We evaluate the success rates of dodging attacks with two methods in the digital and physical world, as shown in Table VII.

In digital scenarios, we use FaceAdv or AGNs to generate stickers and digitally attach them to the 45 face images of each attacker, which results in a total of 900 tests for each target FR system. In physical scenarios, each attacker has 135 images captured from a video clip, and we have 2,700 tests for each target FR system.

[2]We name these combinations from #1 to #10, which consists of taking three from five candidate positions (i.e., right superciliary arch, left superciliary arch, nasal bone, right nasolabial sulcus and left nasolabial sulcus) from left to right in turn.

TABLE VII
SUCCESS RATES (%) IN DODGING OR IMPERSONATING ATTACKS

Mode	Scenario	Method	Target FR System			
			ArcFace	CosFace	FaceNet	VGGFace
Dodging	Digital	FaceAdv	87.33	100.00	100.00	100.00
		AGNs	86.78	94.00	96.00	100.00
	Physical	FaceAdv	77.56	100.00	100.00	100.00
		AGNs	50.78	50.67	62.56	100.00
Impersonating	Digital	FaceAdv	63.44	88.67	94.56	100.00
		AGNs	14.89	48.44	60.33	79.41
	Physical	FaceAdv	8.67	30.89	55.32	62.06
		AGNs	0.00	2.85	5.12	11.98

We have *two key observations* from the attack results.

1) The success rate in the digital world is higher than that in the physical world. When attacking ArcFace in the physical world, the success rate of FaceAdv drops approximately 10%. As for AGNs, the success rate is reduced by 30% on average. This confirms that there is the perturbation loss reduces the effectiveness of adversarial examples.

2) FaceAdv achieves higher success rates in both digital and physical scenarios. Compared with FaceAdv, AGNs has severe perturbation loss in physical scenarios, which results in significant performance degradation (a round 30%) in the physical world. This demonstrates that the sticker generator and the convertor in FaceAdv successfully simulate perturbed images with stickers pasted on real human faces.

2) Impersonating Attacks: For launching impersonating attacks, we employ all the participants in VolFace as attackers and randomly select three victims from LFW^- for each attacker. For each target FR system, there are altogether 2,700 tests in digital scenarios and 8,100 tests in physical scenarios.

The attack success rates are summarized in Table VII. There are *four key observations* from the results.

1) The success rates of FaceAdv are higher than those of AGNs by a large margin in both digital and physical scenarios. It demonstrates the proposed method is superior to AGNs.

2) Compared with the digital world, the success rate of FaceAdv in the physical world reduces by 50% on average. Launching impersonating attacks is more difficult than performing dodging attacks, as the recognition result is the predetermined victim.

3) The performance of FaceAdv on ArcFace is worse than that on the other two systems (i.e., CosFace and FaceNet). As illustrated in Table IV, ArcFace tends to extract facial features from regions other than facial organs, which reduces the effectiveness of crafted stickers and degrades the performance.

4) AGNs can successfully attack FR systems in the digital world but fail in the physical world. The reason is twofold: *first*, the area of eyeglass frames is too small to attack FR systems, as the size of stickers can greatly affect the performance (Section VI-B.2); *second*, the feature extractor is trained on the large-scale dataset (e.g. MS-Celeb-1M [9]) so that it can resist potential adversarial attacks.

VGGFace has been proved to be inferior to the other three FR systems (i.e., ArcFace, CosFace and FaceNet) in the large-scale dataset [9]. Table VII also shows that both AGNs and FaceAdv perform well in VGGFace compared to the other three FR systems in both dodging and impersonating attacks. Thus, we will focus on the other three systems for the evaluation in the following experiments.

A surprising observation is that attackers, when facing directly the camera, can hardly attack ArcFace. This reveals that a normal head pose may not achieve the highest success rate, which motivates us to conduct further investigation on the influence of head poses in Section VI-D.

3) Inconspicuousness of Stickers: Currently, there is no literature that proposes a measure of inconspicuousness of adversarial examples in the physical world. To quantitatively evaluate the inconspicuousness, which is actually a quite subjective characteristic, we borrow the idea of imperceptibility in the digital world and utilize the Euclidean norm (the L_2 norm) [6], the structural similarity index measure (SSIM) [15] and the learned perceptual image patch similarity (LPIPS) [42] based on AlexNet to calculate the difference between the benign and adversarial examples in the physical world. Adversarial examples are facial images with adversarial perturbations (stickers for FaceAdv or eyeglass frames for AGNs). Note that adversarial examples and benign examples are captured under the same environmental conditions, e.g., user-camera distance, brightness and head pose.

Notably, a smaller value of the L_2 norm and LPIPS or a larger value of SSIM indicates that the adversarial example seems more likely to be the original face image and thus achieves better inconspicuousness. In our experiments, the averaged distance values for FaceAdv and AGNs in the L_2 norm, SSIM and LPIPS are 29.81 vs. 36.75, 0.87 vs. 0.84 and 0.21 vs. 0.23, respectively. The results show that the inconspicuousness of FaceAdv is relatively lower than that of AGNs. In the future, we will try to propose more effective approaches for evaluating the inconspicuousness.

D. Evaluation on Influencing Factors

To investigate the performance of FaceAdv in different conditions, we investigate the success rate of FaceAdv by varying several influencing factors, including the user-camera distance, the ambient brightness, and the head pose. In the following experiments, we also use the 20 participants of VolFace as attackers and the 3 individuals from LFW^- as victims for each attacker. For each target FR system, there are 8,100 tests in physical scenarios when evaluating each value taken for each influencing factor.

1) User-Camera Distance: In physical scenarios, attackers (i.e., users) cannot precisely control the distance to the camera, which requires crafted stickers should work with different distances. Thus, we evaluate the success rates of FaceAdv with the user-camera distance of 30cm (e.g. unlocking mobile phone or laptop), 50cm (e.g. passing the access control of buildings) and 70cm (e.g. using face scan payment). The 50cm is the default setting. The results are shown in Table VIII.

From these results, we can find that adversarial stickers crafted by FaceAdv work stably in both digital and physical scenarios. This is mainly because each time when training the generator, FaceAdv attaches adversarial stickers to multiple

TABLE VIII
SUCCESS RATES (%) WITH VARYING DISTANCE

Mode	Target Model	User-Camera Distance (cm)		
		30	50	70
Dodging	ArcFace	77.67	77.56	77.52
	CosFace	100.00	100.00	100.00
	FaceNet	100.00	100.00	100.00
Impersonating	ArcFace	8.70	8.67	8.64
	CosFace	30.91	30.89	30.89
	FaceNet	55.33	55.32	55.30

TABLE IX
SUCCESS RATES (%) WITH VARYING AMBIENT BRIGHTNESS

Mode	Target Model	The Level of Brightness (lux)		
		30	130	250
Dodging	ArcFace	77.70	77.56	77.44
	CosFace	100.00	100.00	100.00
	FaceNet	100.00	100.00	100.00
Impersonating	ArcFace	8.51	8.67	8.89
	CosFace	30.72	30.89	30.70
	FaceNet	55.37	55.32	55.32

face images captured at different distances, as described in Section V-B.

In the physical world, the size of face images is different with varying distances so that the FR system will rescale face images to the resolution of input images. FaceAdv re-samples the final face images with stickers to the certain resolution when training the generator, which ensures stickers are robust with the image re-sampling.

2) Brightness Level: The ambient brightness can also influence the performance of adversarial stickers. We evaluate the success rates of FaceAdv with varying ambient brightness, i.e., 30lux, 130lux (by default) and 250lux. They represent weak indoor light, sunny day, and strong indoor light, respectively. The results are shown in Table IX.

In general, the performance of FaceAdv changes slightly with varying ambient brightness. This is because we employ R-Net to estimate the ambient brightness in Section V-C.2. When digitally attaching crafted stickers to face images, FaceAdv will change the brightness of stickers to fit that of face images according to the parameters of illumination model acquired in advance. The aim of this process is to simulate the appearance of stickers in different ambient brightness so as to improve the robustness of adversarial stickers.

3) Head Pose: As discussed earlier, FaceAdv achieves lower success rate of attacking ArcFace when the participants face the camera directly. We conduct a series of experiments to investigate the success rates with varying head poses.

We select typical head poses: normal (HN, the default head pose), turning head to the left by 20 deg (HR) or to the right by 20 deg (HL), raising up head by 20 deg (HU) or lowering head by 20 deg (HB). The success rates with these head poses are shown in Table X.

FaceAdv achieves the highest success rate (30.06%) on ArcFace when the head pose is HR. In Section VI-B, we select the combination #8 (i.e., left superciliary arch, nasal bone left, nasolabial sulcus) to attach stickers, which achieves the best average performance under different conditions. However, the locations of the combination #8 are mainly on the left side of the face. Thus, when the head pose of participants turns to HR, the stickers attached on the left side are totally exposed to the camera, and the success rate becomes higher than that in the other directions. Apparently, the stickers are basically invisible to the camera in HL, resulting in the worst performance.

TABLE X
SUCCESS RATES (%) WITH VARYING HEAD POSE

Mode	Target Model	Head Pose				
		HN	HL	HR	HU	HB
Dodging	ArcFace	77.56	74.52	85.63	77.26	77.78
	CosFace	100.00	100.00	100.00	100.00	100.00
	FaceNet	100.00	100.00	100.00	100.00	100.00
Impersonating	ArcFace	8.67	3.70	30.06	13.04	8.72
	CosFace	30.89	19.08	34.72	39.65	39.52
	FaceNet	55.32	32.10	33.28	31.31	33.35

TABLE XI
SUCCESS RATES (%) IN THE TRANSFERABILITY OF STICKERS

Mode	Scenario	Method	Target FR System		
			ArcFace	CosFace	FaceNet
Dodging	Digital	FaceAdv-**W**	87.33	100.00	100.00
		FaceAdv-**B**-5	20.11	79.89	81.78
		FaceAdv-**B**-3	10.74	59.96	62.93
	Physical	FaceAdv-**W**	77.56	100.00	100.00
		FaceAdv-**B**-5	15.78	65.22	76.48
		FaceAdv-**B**-3	8.07	53.85	59.07
Impersonating	Digital	FaceAdv-**W**	63.44	88.67	94.56
		FaceAdv-**B**-5	10.89	33.00	54.78
		FaceAdv-**B**-3	1.30	19.03	38.19
	Physical	FaceAdv-**W**	8.67	30.89	55.32
		FaceAdv-**B**-5	5.72	23.53	44.73
		FaceAdv-**B**-3	0.00	18.91	24.47

The performance of FaceAdv is relatively stable for attacking the other two FR systems. This is because we adopt two measures to improve the robustness of FaceAdv with varying head poses. First, we propose the new method based on 3D face reconstruction to digitally attach stickers to face images. Second, each time when training the generator, FaceAdv attaches these stickers onto images captured from different head poses and then feeds them into the target FR system.

E. Evaluation on Transferability of Stickers

The transferability of adversarial stickers means that the stickers crafted against an FR system can cheat the other FR systems, which is critical for black-box scenarios. We utilize the ensemble of source FR systems to train the sticker generator and update the loss function to improve the transferability in Section V-D. In this subsection, we will investigate the transferability of the crafted stickers by FaceAdv.

We employ the 20 participants of VolFace as attackers and the three individuals as victims for each attacker. For attacking a target FR system (e.g., ArcFace), we utilize the other two FR systems (i.e., CosFace and FaceNet) to train the sticker generator. For each target FR system, there are 2,700 tests in digital scenarios and 8,100 tests in physical scenarios.

We also evaluate the influence of the number of stickers on the performance. We regard FaceAdv in white-box scenarios as the baseline (i.e., FaceAdv-**W**), and separately use three (i.e. FaceAdv-**B**-3) and five (i.e., FaceAdv-**B**-5) stickers to attack FR systems for comparison. Since the combination #1 achieves the best performance as shown in Fig. 6 and the target FR system is unknown to the adversary in black-box scenarios, we select the combination #1 for each target FR system in FaceAdv-**B**-3. The results are listed in Table XI.

We have *three key observations* from the results.

1) The transferability of crafted stickers assists FaceAdv in successfully fooling FR systems. Compared with FaceAdv-**W**, the performance of FaceAdv in black-box scenarios (i.e., FaceAdv-**B**-3 and FaceAdv-**B**-5) degrades, but it is still better than that of AGNs in white-box scenarios in some cases. For impersonating attacks against CosFace, the success rate of FaceAdv-**B**-3 is much higher than that of AGNs in white-box scenarios (18.91% vs. 2.85%).

2) The number of stickers is positively related to success rate. For instance, in impersonating attacks against FaceNet, the success rate of FaceAdv-**B**-5 is higher than that of FaceAdv-**B**-3 (44.73% vs. 24.47%). This is because a larger number of stickers generally indicates the larger area of faces to be perturbed, making the perturbations more effective.

3) In these three FR systems, FaceNet is more vulnerable to the crafted stickers, while ArcFace is more resistant. As illustrated in Table IV, FaceNet and CosFace mainly focus on facial organs to extract information while ArcFace does not. Due to this difference, when using CosFace and FaceNet to train the sticker generator, the performance of FaceAdv on ArcFace drops a lot, especially in the physical world.

VII. Discussion

The results of our study show that FaceAdv is an effective algorithm for generating physical-world adversarial examples against the state-of-the-art FR systems. Now, we discuss the limitations of FaceAdv as well as the future directions.

A. Liveness Detection

FaceAdv can limit the size and the shape of crafted stickers, which can help the attacker pass the liveness detection. The detection systems always employ the eye and mouth movement [31] for the presentation attack detection. The stickers are relatively small and do not block the eyes so that these systems will not alert.

B. Face Detection

The adversarial stickers will not influence the performance of face detector. When training the sticker generator, the feature extractor and the MLP classifier are applied to produce the recognition result. Therefore, the face detector is unknown to the crafted stickers. In our experiments, the percentage of the frames in which faces cannot be detected is less than 0.01%.

C. Sticker Size

The size and location of printed stickers will affect the performance. Ideally, stickers should be the same as those in the digital world. However, face images are captured in different directions and the sizes of human faces are also different, making the size of printed stickers hardly estimated. We refer to this gap as *fabrication error*.

We randomly re-sample stickers before feeding them into the convertor to simulate the difference of sticker size and the location between digital and physical scenarios. Specifically, we stochastically scale (0.9 to 1.1), rotate ($-$10deg to 10deg) and translate ($-$10px to 10px) stickers to imitate the error, which improves the robustness of FaceAdv.

D. Convertor

The performance of the convertor determines the difference of face images with stickers between digital scenarios and physical scenarios. A more effective algorithm for 3D face reconstruction improves the performance of FaceAdv.

An alternative choice is to abandon the convertor and directly place stickers on face images. In that case, the renderer calculates the sticker template in reverse and cuts physical stickers out. However, the shape of physical stickers is irregular because of the radian of face, which enhances the difficulty of tailoring stickers. If the convertor is removed, the fabrication error will cause the significant performance degradation. Therefore, we use the methodology described in Fig. 3 rather than this alternative choice.

E. Black-Box Scenarios

As shown in Table XI, the performance of FaceAdv for attacking ArcFace in black-box scenarios is relatively poor by simply transferring from the other FR systems. Critical regions where ArcFace extracts information are different with the other two FR systems, as illustrated in Table IV, which leads to performance degradation.

In order to improve the performance, when FaceAdv optimizes parameters of the sticker generator, the gradient of face images with stickers must be estimated to point out the direction of optimization. In the future work, we can utilize the Monte Carlo gradient estimation and the prior-guided random gradient-free method [2] to estimate the gradient.

F. Countermeasures

There are two possible defenses against FaceAdv, i.e., adversarial training [25] and sticker detection [16]. Adversarial training, in which a network is trained on facial images with crafted stickers, is a general approach to defend against adversarial attacks. Unfortunately, FaceAdv spends 20 minutes to train the sticker generator for each attacker in dodging attacks or each attacker-victim pair in impersonating attacks, which makes adversarial training impractical in large-scale datasets. The core idea of sticker detection is that, when removing the critical regions from the input images, the ranking changes of adversarial labels will be larger than those of benign labels. However, FaceAdv manipulates several regions at the same time and removing one region that cannot make significant ranking changes of adversarial labels. We will explore effective countermeasures in the future work.

VIII. Conclusion

In this paper, we proposed a method named FaceAdv to automatically generate adversarial stickers, misleading the results of FR systems in the physical world. We employed an architecture of GANs to train a generator to craft adversarial stickers so as to fabricate a large number of stickers with different shapes after training, and proposed a novel method RSO to digitally attach these stickers to face images. Extensive experimental results demonstrated that FaceAdv can achieve high success rate in physical scenarios with different environmental conditions. In future work, we will further investigate techniques to improve the effectiveness of FaceAdv and explore powerful defenses.

References

[1] A. J. Bose and P. Aarabi, "Adversarial attacks on face detectors using neural net based constrained optimization," in *Proc. 20th IEEE Int. Workshop Multimedia Signal Process. (MMSP)*, Vancouver, BC, Canada, Aug. 2018, pp. 1–6.

[2] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., Vancouver, BC, Canada, Dec. 2019, pp. 10932–10942.

[3] A. Dabouei, S. Soleymani, J. M. Dawson, and N. M. Nasrabadi, "Fast geometrically-perturbed adversarial faces," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Waikoloa Village, HI, USA, Jan. 2019, pp. 1979–1988.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 4690–4699.

[5] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 285–295.

[6] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1625–1634.

[7] G. Garofalo, V. Rimmer, T. V. hamme, D. Preuveneers, and W. Joosen, "Fishy faces: Crafting adversarial images to poison face authentication," in *Proc. 12th USENIX Workshop Offensive Technol. (WOOT)*, C. Rossow and Y. Younan, Eds., Baltimore, MD, USA, Aug. 2018, pp. 1–12.

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., Long Beach, CA, USA, Dec. 2017, pp. 5767–5777.

[9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Computer Vision* (Lecture Notes in Computer Science), vol. 9907, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 87–102.

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

[11] E. Kaziakhmedov, K. Kireev, G. Melnikov, M. Pautov, and A. Petiushko, "Real-world attack on MTCNN face detection system," in *Proc. Int. Multi-Conf. Eng., Comput. Inf. Sci. (SIBIRCON)*, Oct. 2019, pp. 422–427.

[12] M. Killioglu, M. Taskiran, and N. Kahraman, "Anti-spoofing in face recognition with liveness detection using pupil tracking," in *Proc. IEEE 15th Int. Symp. Appl. Mach. Intell. Informat. (SAMI)*, Jan. 2017, pp. 87–92.

[13] 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.

[14] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey," *Sensors*, vol. 20, no. 2, p. 342, Jan. 2020.

[15] C. Laidlaw, S. Singla, and S. Feizi, "Perceptual adversarial robustness: Defense against unseen threat models," in *Proc. 9th Int. Conf. Learn. Represent.*, Virtual Event, Austria, May 2021, pp. 1–25.

[16] F. Li, X. Liu, X. Zhang, Q. Li, K. Sun, and K. Li, "Detecting localized adversarial examples: A generic approach using critical region analysis," *CoRR*, vol. abs/2102.05241, pp. 1–11, Feb. 2021.

[17] Linda. (2017). *Unattended Convenience Store, New Trends in Retail Industry*. Accessed: Oct. 9, 2020. [Online]. Available: https://www.lsvisionhd.com/news/Unattended-Convenience-Store-New-Trends-in-Retail-Industry.html

[18] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. 5th Int. Conf. Learn. Represent.*, Toulon, France, Apr. 2017, pp. 1–24.

[19] Megvii. (2020). *Smart Building Access Solution*. Accessed: Oct. 9, 2020. [Online]. Available: https://en.megvii.com/solutions/Access_Control_for_Smart_Building

[20] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. Asia Conf. Comput. Commun. Secur.*, R. Karri, O. Sinanoglu, A. Sadeghi, X. Yi, Eds., Abu Dhabi, United Arab Emirates, Apr. 2017, pp. 506–519.

[21] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, X. Xie, M. W. Jones, G. K. L. Tam, Eds., Swansea, U.K., Sep. 2015, pp. 41.1–41.12.

[22] M. Pautov, G. Melnikov, E. Kaziakhmedov, K. Kireev, and A. Petiushko, "On adversarial patches: Real-world attack on ArcFace-100 face recognition system," in *Proc. Int. Multi-Conf. Eng., Comput. Inf. Sci. (SIBIRCON)*, Oct. 2019, pp. 0391–0396.

[23] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 815–823.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[25] A. Shafahi *et al.*, "Adversarial training for free!" in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., Vancouver, BC, Canada, Dec. 2019, pp. 3353–3364.

[26] 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. Conf. Comput. Commun. Secur.*, E. R. Weippl, S. Katzenbeisser, C. Kruegel, A. C. Myers, and S. Halevi, Eds., Vienna, Austria, Oct. 2016, pp. 1528–1540.

[27] 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, Jul. 2019.

[28] M. Shen, Z. Liao, L. Zhu, K. Xu, and X. Du, "VLA: A practical visible light-based attack on face recognition systems in physical world," *ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–19, Sep. 2019.

[29] M. Shen, Y. Liu, L. Zhu, X. Du, and J. Hu, "Fine-grained webpage fingerprinting using only packet length information of encrypted traffic," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2046–2059, 2021.

[30] M. Shen *et al.*, "Exploiting unintended property leakage in blockchain-assisted federated learning for intelligent edge computing," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2265–2275, Feb. 2021.

[31] M. Shen, Y. Wei, Z. Liao, and L. Zhu, "IriTrack: Face presentation attack detection using iris tracking," *ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–21, Jun. 2021.

[32] M. Shen, J. Zhang, L. Zhu, K. Xu, and X. Du, "Accurate decentralized application identification via encrypted traffic analysis using graph neural networks," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2367–2380, 2021.

[33] Vee Technologies. (2019). *Face Recognition in Cars Improves Safety and Convenience*. Accessed: Oct. 9, 2020. [Online]. Available: https://visagetechnologies.com/face-recognition-in-cars/

[34] H. M. Tummon, J. Allen, and M. Bindemann, "Facial identification at a virtual reality airport," *I-Perception*, vol. 10, no. 4, 2019, Art. no. 2041669519863077.

[35] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5265–5274.

[36] A. Wójtowicz and J. Chmielewski, "Face-based passive customer identification combined with multimodal context-aware payment authorization: Evaluation at point of sale," in *Proc. 20th Int. Conf. Enterprise Inf. Syst.*, vol. 1, S. Hammoudi, M. Smialek, O. Camp, J. Filipe, Eds. Funchal, Portugal: SciTePress, 2018, pp. 555–566.
[37] 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.
[38] X. Yang, W. Liu, S. Zhang, W. Liu, and D. Tao, "Targeted attention attack on deep learning models in road sign recognition," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4980–4990, Mar. 2021.
[39] X. Yang, F. Wei, H. Zhang, and J. Zhu, "Design and interpretation of universal adversarial patches in face detection," in *Computer Vision* (Lecture Notes in Computer Science), vol. 12362, A. Vedaldi, H. Bischof, T. Brox, J. Frahm, Eds. Glasgow, U.K.: Springer, Aug. 2020, pp. 174–191.
[40] 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.
[41] L. Zhang, Y. Meng, J. Yu, C. Xiang, B. Falk, and H. Zhu, "Voiceprint mimicry attack towards speaker verification system in smart home," in *IEEE IEEE Conf. Comput. Commun.*, Jul. 2020, pp. 377–386.
[42] 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.*, Salt Lake City, UT, USA, Jun. 2018, pp. 586–595.
[43] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang, "Invisible mask: Practical attacks on face recognition with infrared," *CoRR*, vol. abs/1803.04683, pp. 1–13, Mar. 2018.

Meng Shen (Member, IEEE) received the B.Eng. degree in computer science from Shandong University, Jinan, China, in 2009, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2014. He is currently an Associate Professor with Beijing Institute of Technology, Beijing. He has authored over 50 papers in top-level journals and conferences, such as ACM SIGCOMM, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC), and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (TIFS). His research interests include data privacy and security, blockchain applications, and encrypted traffic classification. He received the Best Paper Award from IEEE/ACM IWQoS 2021. He was selected by the Beijing Nova Program 2020 and the winner of the ACM SIGCOMM China Rising Star Award in 2019. He has guest edited Special Issues on Emerging Technologies for Data Security and Privacy in *IEEE Network* and IEEE INTERNET OF THINGS JOURNAL.

Hao Yu received the B.Eng. degree in computer science from Inner Mongolia University, Hohhot, China, in 2019. He is currently pursuing the master's degree with the School of Computer Science, Beijing Institute of Technology. His research interest includes adversarial example.

Liehuang Zhu (Member, IEEE) is currently a Professor with the Department of Computer Science, Beijing Institute of Technology. He is selected into the Program for New Century Excellent Talents in University from Ministry of Education, China. His research interests include the Internet of Things, cloud computing security, internet, and mobile security.

Ke Xu (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China. He currently serves as a Full Professor for Tsinghua University. He has published more than 200 technical articles and holds 11 U.S. patents in the research areas of next-generation internet, blockchain systems, the Internet of Things, and network security. He is a member of ACM. He served as the Steering Committee Chair for IEEE/ACM IWQoS. He is an Editor of IEEE INTERNET OF THINGS JOURNAL. He has guest edited several special issues in IEEE and Springer journals.

Qi Li (Senior Member, IEEE) received the Ph.D. degree from Tsinghua University. He has worked with ETH Zürich and The University of Texas at San Antonio. He is currently an Associate Professor with the Institute for Network Sciences and Cyberspace, Tsinghua University. His research interests include network and system security, particularly in internet and cloud security, mobile security, and big data security. He is currently an Editorial Board Member of the IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING (TDSC) and ACM DTRAP.

Jiankun Hu (Senior Member, IEEE) is currently a Professor with the School of Engineering and Information Technology, University of New South Wales, Canberra, ACT, Australia. He has many publications in top venues, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, *Pattern Recognition*, and IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS. He is an Invited Expert of Australia Attorney-General's Office assisting the draft of Australia National Identity Management Policy. His research interests include cyber security covering intrusion detection, sensor key management, and biometrics authentication. He has served at the Panel on Mathematics, Information and Computing Sciences, Australian Research Council, Excellence in Research for Australia (ERA) Evaluation Committee, in 2012. He is an Associate Editor of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.