This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

Adversarial Attacks against Face Recognition: A Comprehensive Study

Fatemeh Vakhshiteh[1], Ahmad Nickabadi[2], and Raghavendra Ramachandra[3]
[1]Department of Biomedical Engineering, Amirkabir University of Technology, Tehran, Iran
[2]Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran
[3]Department of Information Security and Communication Technology, Norwegian Biometrics Laboratory (NBL), Norwegian University of Science and Technology (NTNU i Gjovik), Gjøvik, Norway

Corresponding author: Raghavendra Ramachandra (e-mail: raghavendra.ramachandra@ntnu.no).

ABSTRACT Face recognition (FR) systems have demonstrated reliable verification performance, suggesting suitability for real-world applications ranging from photo tagging in social media to automated border control (ABC). In an advanced FR system with deep learning-based architecture, however, promoting the recognition efficiency alone is not sufficient, and the system should also withstand potential kinds of attacks. Recent studies show that (deep) FR systems exhibit an intriguing vulnerability to imperceptible or perceptible but natural-looking adversarial input images that drive the model to incorrect output predictions. In this article, we present a comprehensive survey on adversarial attacks against FR systems and elaborate on the competence of new countermeasures against them. Further, we propose a taxonomy of existing attack and defense methods based on different criteria. We compare attack methods on the orientation, evaluation process, and attributes, and defense approaches on the category. Finally, we discuss the challenges and potential research direction.

INDEX TERMS Biometrics, Face recognition, Adversarial attacks, Adversarial perturbation, Deep learning,

I. INTRODUCTION

Face recognition (FR) has been a prevalent biometric technique for identity authentication and is broadly used in several areas, such as finance, military, public security, and daily life. A typical FR system's ultimate goal is to identify or verify a person from a digital image or a video frame. Researchers describe FR as a biometric artificial intelligence-based application that can exclusively identify a person through analyzing patterns of the person's facial features.

The idea of using the face as a biometric trait inspired in the 1960s, and the design of the first successful FR system dates back to the early 1960s [1]. In recent times, the latest advancements of deep learning and the use of mounting hardware and abundant data have resulted in massive development in FR algorithms with accurate performance [2]–[4]. This performance permits the broad deployment of FR technologies in further diverse applications, ranging from photo tagging in social media to dubious identification in automated border control (ABC) systems.

In an advanced FR model, however, promoting the recognition efficiency alone is not sufficient, and the system should also withstand potential kinds of attacks. Recently, researchers found that (deep) FR systems are vulnerable against different types of attacks that create data variations to fool classifiers. These attacks can be launched either via (a) physical attacks, which modify the physical appearance of a face before image capturing, or (b) digital attacks, which implement modifications in the captured face image [5].

Presentation attacks also referred to as spoofing attacks [6], are among the main techniques used for physical attacks. A presentation attack aims to subvert the face recognition system by presenting a facial biometric artifact, including a printed photo, the electronic display of a facial photo, replaying video using an electronic display, and 3D face masks [7]. It has recently been demonstrated that makeup can also be abused to launch presentation attacks [8].

In contrast, adversarial attacks [9] and the variations resulting from morphing attacks [10, 116] are critical techniques utilized for digital invasion. A typical adversarial attack can deceive the FR systems with carefully crafted perturbations, called adversarial examples [11]. It should be noted that adversarial attacks are mainly categorized in the class of digital attacks, e.g., adversarial example generation methods mostly implement on the face images digitally, however, some methods are designed to accomplish physically by making physical changes on the face

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

appearances and then capturing the modified images [12]. Several approaches have been proposed to overcome the devastating consequences of this type of attacks, both those that target FR systems [13]–[16] and those that target beyond that area [17]–[19]. On the other hand, the goal of a morphing attack is to generate a fake face with the morphing and blending of two or more different subjects (e.g., a criminal and an accomplice) to enroll the criminal as a legitimate identity template of the FR system [20], [21]. Similarly, many efforts have been made in this regard to countermeasure destructive consequences ranging from face morphing detection methods [22]–[25] to accomplice's facial restoration approaches [26]–[28].

Among different attacks, adversarial attacks are fascinating since they can generally target deep neural networks (DNNs) and could specifically focus on convolutional neural networks (CNNs), based on which the state-of-the-art FR models are established. The massive growth in the number of papers published each year in the field of adversarial example generation demonstrates this type of attack (see Fig. 1).

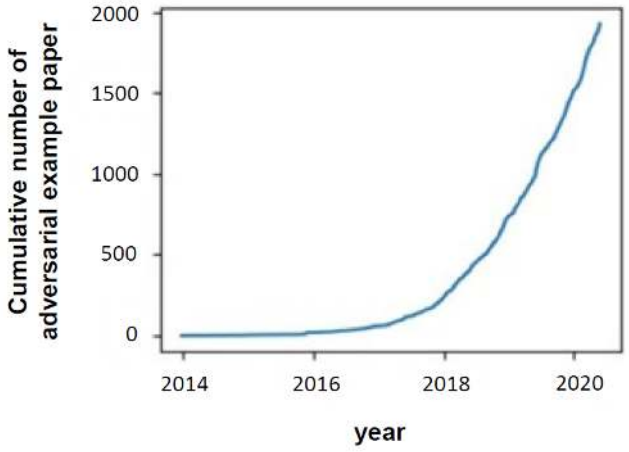

FIGURE 1. The cumulative number of adversarial example papers published in recent years [29].

This work presents a comprehensive survey on different techniques of adversarial attack generation intended to deceive FR systems, along with the potential countermeasures established against them. This is the first study that attempts to review adversarial attack and defense strategies on FR systems to the best of our knowledge. Since FR may refer to each of the two applications of face identification or face verification, we review both in this study.

The main contributions of this paper are:

- We review recent studies on adversarial example generation approaches on FR systems, present an illustrative taxonomy of the corresponding methods according to their orientation, and compare these approaches on orientation, evaluation process, and attributes.
- We review the new adversarial detection methods for the FR systems, categorize the presented algorithms, and demonstrate a descriptive taxonomy.
- We outline the main challenges and potential solutions for adversarial examples that target FR models based on four main problems: Particularization/Specification of adversarial examples, instability of FR models, deviation from the human vision system, and image-agnostic perturbation generation.

The remainder of this paper is organized as follows: Section II introduces the background of FR techniques, architectures, and datasets. In Section III, we describe the standard terms related to adversarial attacks and defenses in the context of the FR course, represent the attacks' attributes, explain the experimental standards, and discuss the pioneer methods of generating attacks. We review adversarial example generation methods intended to deceive the FR mission in Section IV. We discuss the methods and compare the approaches based on orientation, evaluation process, and attributes. In Section V, corresponding countermeasures are investigated. We discuss current challenges and potential future research directions in Section VI. Section VII concludes the work.

II. BACKGROUNDS

In this section, we briefly introduce basic FR systems and elaborate on incorporated models in the era of deep learning. Next, we present widely used architectures and standard datasets in this regard.

A. A BRIEF INTRODUCTION TO FACE RECOGNITION

Face recognition has been an age-old research topic in the computer vision community, and the first success of it dates to the 1960s. Since then, this research path has undergone scientific leaps in four decisive times. The face representation for recognition has taken sequential forms of holistic learning, local feature learning, shallow learning, and deep learning [30].

In the early 1990s, the historical Eigenface approach [1] was introduced, and the study of FR became popular shortly after that. From then till the 2000s, the holistic approaches that extracted low-dimensional representations from face images based on certain distribution assumptions [31]–[34] dominated the FR community. Nevertheless, these methods demonstrated a failure in addressing the uncontrolled facial modifications that deviate from the prior considered assumptions. In the early 2000s, local-feature-based FR techniques were introduced, and handcrafted descriptors such as Gabor [35] and LBP [36] became popular. However, distinctiveness and compactness were the two properties these local features lacked. In the early 2010s, local learning-based features were introduced [37]–[39] to learn local filters and encode codebooks for better distinctiveness and compactness Though resolved the lack of necessary properties, these shallow representations demonstrated a loss

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

of robustness against complicated nonlinear facial appearance variations.

These traditional methods attempted to recognize faces by one- or two-layer representations and improved FR accuracy The goal is to explore each aspect of unconstrained facial variations, including illumination, pose, expression, or occlusion, separately. The advent of deep learning methods resolved the limitations of traditional methods. In deep-learning-based FR approaches, multiple layers of processing units learn multiple representations that correspond to different levels of abstraction. Interestingly, the higher-level abstract representations have demonstrated a strong invariance against face illumination, pose, expression, and occlusion changes, and represented facial identity with extraordinary stability. In 2014, DeepFace [3] attained state-of-the-art accuracy on the Labeled Faces in the Wild (LFW) dataset [40]. In an unconstrained condition, it competed successfully with the human performance for the first time and approached the desired accuracy by training a 9-layer network on 4 million facial images. Deep learning techniques have reformed the research horizon of FR in almost all aspects, from algorithm designs and training/test datasets to application setups and evaluation protocols.

B. DISTINGUISHED ARCHITECTURES OF FACE RECOGNIZERS

DeepFace [3] was the first distinguished deep architecture introduced to the FR community. It has a deep CNN architecture with several locally connected layers. Afterward, FaceNet [41] and VGG-Face [2] deep-learning-based models were introduced, which were designed to train popular GoogleNet [42] and VGGNet [43] over the large-scale face datasets, respectively. These models fine-tuned the networks via a triplet loss function and implemented it on face patches created by an online triplet mining method. Later, the SphereFace [44] was proposed according to ResNet architecture [45], and a novel angular softmax loss learns discriminative features by an angular margin. Similar to this network, CosFace [46] and ArcFace [47] were introduced based on cosine and angular margin-based loss, respectively. These models were designed in a way to separate learned features with larger cosine and angular distances. Lightweight networks were then proposed to overcome the lack of GPUs' power and memory size and become applicable to many mobiles and embedded devices. LightCNN [48], with a novel max-feature-map (MFM) activation function, is a famous example of this category that results in a compact representation and reduces the computational cost.

C. STANDARD FACE RECOGNITION DATASETS

In 2007, the LFW dataset was provided from 3K images of faces on the web under unconstrained conditions and opened a new path for other testing databases to be used in different tasks. Having sufficiently large training datasets to evaluate the effectiveness of deep FR models resulted in continually developing more complex datasets to facilitate the FR research. The early deep FR models, such as DeepFace, FaceNet, and DeepID [49], were trained on private, controlled, or small-scale training datasets, hence, not allowing the new models to compare with. To resolve this problem, CASIA-Webface [50], a collection of 0.5M images of 10K celebrities, was introduced as the first widely used public training dataset. Later, MS-Celeb-1M [51], VGGface2 [52], and Megaface [53], collections of over 1M images, were introduced as a public large-scale training dataset to be used by many advanced deep learning methods.

III. ADVERSARIAL ATTACK GENERATION

An adversarial attack consists of finely modifying an original image with the intention of the alterations become almost imperceptible to the human eye, to fool a specific classifier. In the realm of digital attacks, this can be implemented as the addition of a minimal vector $\boldsymbol{n}$ to the input image $\boldsymbol{x}$, i.e. ($\boldsymbol{x}$ + $\boldsymbol{n}$), such that the deep learning model $\boldsymbol{F}$ predicts an incorrect output for the altered input $\boldsymbol{x}$ + $\boldsymbol{n}$, which is known as an adversarial example. This way, a box-constrained optimization problem for generating the adversarial example $\boldsymbol{x}'$ can generally be described as [9]:

$$\begin{aligned} &\min_{x'} \ \|\boldsymbol{x}' - \boldsymbol{x}\|_2 \\ &s.t. \ \ \boldsymbol{F}(\boldsymbol{x}') = l' \\ &\boldsymbol{F}(\boldsymbol{x}) = l \\ &l \neq l' \\ &\boldsymbol{x}' \in [0,1] \end{aligned} \tag{1}$$

where l and l' represent the output label of $\boldsymbol{x}$ and $\boldsymbol{x}'$, and $\|.\|_2$ denotes the distance between two image samples according to L_2-norm.

As represented in Fig. 2, to fool the FR model (VGG16 in this case), the input images are altered so that the human can still forecast the correct class. However, deep learning network will be confused and misled to the wrong category. Szegedy *et al.* [9] were the first to demonstrate the vulnerability of CNN models to adversarial attacks generated by introducing a minute noise in the input image. The accuracies of GoogleNet and VGG-Face models also demonstrated to be degraded with color balance manipulation. Note that adversarial attacks' invisibility and the widespread application of deep learning algorithms can cause severe damages in real-world scenarios [54]. For example, if the signboard is altered in self-directed driving, adversarial examples can overly threaten the car, pedestrians, and other automobiles. Similarly, in FR applications, the failure to verify the altered input could lead to the degraded performance that can take benefit in the closed set verification/identification scenarios.

A. TERMS AND DEFINITIONS

This section gives a brief introduction to the standard terms related to adversarial attacks on (deep) FR models. Our definitions of words are essential to understand the technical

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

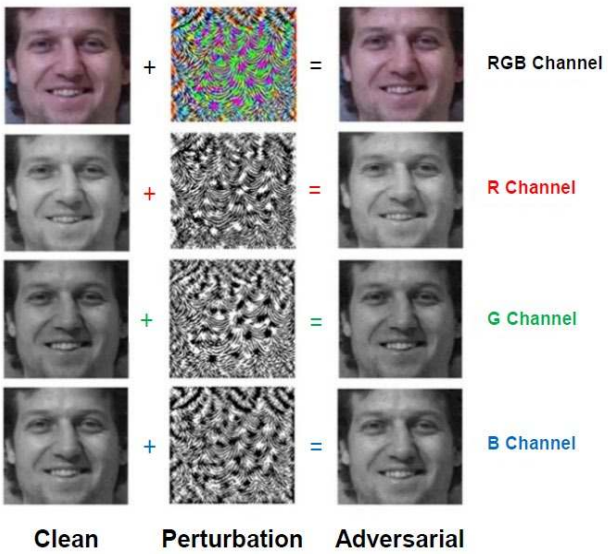

FIGURE 2. Visualization of original face image (first column), adversarial noise vector of VGG-16 (second column), and altered image (last column). From top to bottom, the four rows represent the addition of adversarial noise to the original RGB image and corresponding grayscale representations of R, G, and B color *channels*. Adversarial noise is magnified by a factor of 4 to enhance visibility [13].

components of the reviewed studies. The remainder of this article follows the same definitions of the terms.

- ***Adversarial example/image*** is an intentionally altered (e.g., by adding noise) version of a clean image to fool machine learning (ML) models, such as FR models.
- ***Adversarial training*** is a training process that uses adversarial images along with clean images.
- ***Adversary*** is an agent who creates an adversarial example or the example itself, depending on the case study.
- ***Dodging attack*** occurs when the attacker tries to have a face misidentified as any other arbitrary face. It is also known as *obfuscation attack* in the literature [55], [56].
- ***Evasion attack*** tries to evade the system by altering samples during the testing phase yet not influencing the training data.
- ***Impersonation attack*** seeks to disguise a face as a specific (authorized) face.
- ***Poisoning attack*** takes place during the training time to contaminate the training data. In this attack, the attacker tries to poison data by inserting wisely designed samples to compromise the whole learning process ultimately.
- ***Threat model*** is a model that formalizes assumptions about the attacker's goals, attack strategy, knowledge of the attacked system.

B. ADVERSARIAL ATTACKS ATTRIBUTES

In this section, we discuss the main attributes of adversarial example generation methods.

1) ADVERSARIAL CAPACITY

Adversarial capacity is determined by the amount of knowledge the attackers could gain about the model. Threat models in deep FR systems are classified into the following types according to the attack's capacity.

White-box attack assumes the complete knowledge of the target model, i.e., its parameters, architecture, training method, and even in some cases, its training data.

Black-box attack feeds a target model with the adversarial examples (during testing) created without knowing that model (e.g., its training procedure or its architecture or its parameters). Though the knowledge of the model is not available, the attackers can interact with such a model by utilizing the transferability of adversarial examples (Section III-B.3).

2) ADVERSARIAL SPECIFICITY

Adversarial specificity is defined as the ability of the attack to allow a specific intrusion/disruption or create general mayhem. Threat models in deep FR systems could be categorized into the following types according to the attack's specificity.

Targeted attack deceives a model into falsely predicting a specific label for the adversarial example. In an FR or biometric system, this is achieved by impersonating distinguished people.

Non-targeted attack predicts the adversarial examples' labels irrelevantly, as long as the results are not the correct labels. In an FR/biometric system, this is accomplished through face dodging. A non-targeted attack is more comfortable to implement than a targeted attack since it has more choices and space to alter the output.

3) ADVERSARIAL TRANSFERABILITY

Adversarial transferability is the ability of an adversarial example to continue to impact the models other than the one employed to create it. It is critical for black-box attacks where access to the target model, the training dataset, and other learning parameters may not be available. A substitute neural network model can be trained in such circumstances, and then adversarial examples can be generated against the substitute model. Due to transferability, the target model will be vulnerable to these adversarial examples. The transferability of adversarial examples could be defined from easy to hard, according to having the same neural network architectures but different datasets or having different neural network architectures from the beginning [11].

4) ADVERSARIAL PERTURBATIONS

Adversarial perturbation is a kind of disruption that can fool a given model on a specific image with high probability. Small perturbation is a central premise for adversarial examples. In the realm of adversarial machine learning, the goal is to minimize the norm of the smallest adversarial perturbation to make target models misclassified. Explicitly, given an input image $\boldsymbol{x}$, the perturbation vector $\boldsymbol{n}$ aims to alter the label of $\boldsymbol{x}$, corresponding to the minimal distance from $\boldsymbol{x}$ to the decision boundary of the classifier [9]:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

$$\min_{\boldsymbol{n}\in\mathbb{R}^d} \|\boldsymbol{n}\|_2 \quad s.t. \quad \boldsymbol{F}(\boldsymbol{x})\boldsymbol{F}(\boldsymbol{x}+\boldsymbol{n}) \leq 0 \tag{2}$$

where d is the dimension of the input image and perturbation vector. The perturbation could be categorized into the following types according to the scope of its implementation.

Image-specific perturbations can be explicitly generated according to the given input images.

Universal perturbations can be generated without knowing the underlying details of the given images. Note that universality refers to the characteristic of a perturbation to have a good transferability and the ability to be applied to all input data uniformly. Although universal perturbations make it easier to create adversaries in real-world applications, most present attacks generate image-specific perturbations. It is aimed to move toward this direction and create universal perturbations that are not required to be reformed when the input samples are changed (Section VI).

C. EXPERIMENTAL STANDARDS

The performance of adversarial attacks against FR systems is evaluated based on different datasets and target models. This spectrum results in complications to evaluate the adversarial attacks and quantify the robustness of FR models. Large datasets and complex models usually make the attack and defense exertions harder.

Datasets. The LFW, CASIA-WebFace, MegaFace, VGGFace2, and CelebA [57] are the most widely used image classification datasets to evaluate adversarial attacks on FR systems.

Target models. Adversaries broadly attack several eminent deep FR models, such as DeepFace, FaceNet, VGG-Face, DeepID, SphereFace, CosFace ArcFace, OpenFace [58], dlib[1] , and LResNet100E-IR Face ID model[2].

According to these datasets and target models in the following sections, we will inspect recent studies on adversarial examples targeted FR models according to these datasets and target models.

D. PIONEERS

In this section, we review several pioneer methods for generating adversarial examples. Almost each one of these methods forms the basis of the real-world attacks and has the power of significantly affecting machine learning target models in practice. Descriptions provided here will show the gradual improvements of the adversarial attacks and the extent to which state-of-the-art adversarial attacks can achieve. We will focus on the main methods that attack DNNs in general and review them in chronological order to maintain discussion flow.

1) L-BFGS

Szegedy *et al.* [9] first generated adversarial examples using an *L-BFGS* method. The box-constrained *L-BFGS* is used for approximately solving the following problem:

$$\min_{\boldsymbol{x}} \; c\|\boldsymbol{n}\|_2 + L(\boldsymbol{x}', l) \quad s.t. \quad \boldsymbol{x}' \in [0,1] \tag{3}$$

where $L(\boldsymbol{x}', \boldsymbol{l})$ computes the loss of the classifier, and a minimum $c > 0$ is approximately calculated by line-searching to satisfy the above condition. Authors showed that the above method could compute perturbations that fool neural networks when added to clean images while remains imperceptible to human eyes.

2) FAST GRADIENT SIGN METHOD (FGSM)

Goodfellow *et al.* [59] proposed a fast and straightforward method, named *Fast Gradient Sign Method (FGSM)* to compute an adversarial perturbation by solving the following problem efficiently:

$$\boldsymbol{n} = \epsilon \, sign\big(\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{\theta}, \boldsymbol{x}, l)\big) \tag{4}$$

where ϵ is the perturbation magnitude, $\mathrm{sign}(.)$ denotes the sign function, and $\nabla_{\boldsymbol{x}}\mathcal{J}(.,.,.)$ represents the gradient of the cost function around the current value of the model parameters concerning the $\boldsymbol{x}$. The generated adversarial example $\boldsymbol{x}'$ is calculated as $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{n}$. With the application of the *FGSM* method, adversarial examples are not computed iteratively but, in a one-step, gradient update along the direction of the gradient sign at each pixel. Miyato *et al.* [60] proposed a closely related method and named it *Fast Gradient* L_2. With this method, the perturbation is computed as:

$$\boldsymbol{n} = \epsilon \frac{\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{\theta}, \boldsymbol{x}, l)}{\|\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{\theta}, \boldsymbol{x}, l)\|_2} \tag{5}$$

As it is shown, the computed gradient is normalized with its L_2-norm. An alternative of using the L_∞-norm for normalization was proposed by Kurakin *et al.* [61] and referred to as the *Fast Gradient* L_∞ method. In the literature, all of these methods are categorized as one-step methods.

3) BASIC & LEAST-LIKELY ITERATIVE CLASS METHODS

Kurakin *et al.* [54] extended the one-step gradient ascent idea and proposed the *Basic Iterative Method (BIM)*. The *BIM* iteratively adjusts the direction that increases the loss of the classifier by running multiple small steps. In each iteration, the values of the pixels of the image are clipped as follows:

$$\boldsymbol{x}'^{(i+1)} = Clip_{\epsilon}\Big\{\boldsymbol{x}'^{(i)} + \alpha \cdot sign\Big(\nabla_{\boldsymbol{x}'^{(i)}} \mathcal{J}\big(\boldsymbol{\theta}, \boldsymbol{x}'^{(i)}, l\big)\Big)\Big\} \tag{6}$$

[1] http://dlib.net

[2] https://github.com/deepinsight/insightface/wiki/Model-Zoo

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

where $\boldsymbol{x}^{'(i)}$ denotes the generated adversarial example at the i^{th} iteration, $Clip_{\epsilon}\{.\}$ confines its change in each iteration, and α is the step size. The initialization of the *BIM* algorithm is done by setting $\boldsymbol{x}^{'(0)} = \boldsymbol{x}$, and its termination is controlled by the number of iterations determined by $min(\epsilon + 4,\ 1.25\epsilon)$. This method is also known as the *Iterative Fast Gradient Sign Method (I-FGSM)* in the literature. Following this methodology, the *Iterative Fast Gradient Value Method (I-FGVM)* is proposed, which differs in how it uses the $\nabla_{x'(i)}\mathcal{J}$ gradient [54], [62]. Specifically, the *I-FGVM* changes the input $\boldsymbol{x}$ in the direction of the gradient, whereas the *I-FGSM* uses only the sign gradient. In each iteration of *I-FGSM*, the values of the pixels of the image are clipped as follows:

$$\boldsymbol{x}^{'(i+1)} = Clip_{\epsilon}\{\boldsymbol{x}^{'(i)} + \alpha \cdot \nabla_{\boldsymbol{x}^{'(i)}}\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{x}^{'(i)}, l)\} \quad (7)$$

In another try, Kurakin *et al.* [54] extended *BIM* to *Iterative Least-likely Class Method (ILCM)*, similar to what they did to extend *FGSM* to its "one-step target class." They substituted the label l of the image in (6) by the least likely class (say l_2) predicted by the classifier and tried to maximize the cross-entropy loss.

4) JACOBIAN-BASED SALIENCY MAP ATTACK (JSMA)

Papernot *et al.* [63] designed an adversarial attack by confining the L_0-norm of the perturbations. In contrast to perturbing the whole image, they planned to perturb a few pixels in the image that might induce significant changes to the output. Accordingly, they defined a saliency adversarial map, called *Jacobian-based Saliency Map Attack (JSMA)*, by which they could monitor the effect of changing each pixel of the clean image on the resulting classification. The proposed algorithm is repeated until the maximum number of allowable pixels are altered in the adversarial image so that the neural network fooling succeeded.

5) ONE PIXEL ATTACK

Su *et al.* [64] proposed a successful method of fooling different neural networks by only changing one pixel per image. The optimization problem becomes:

$$\begin{aligned} &\min_{\boldsymbol{x}'} \mathcal{J}(\boldsymbol{\theta}, \boldsymbol{F}(\boldsymbol{x}'),\ l') \\ &s.t.\ \ \|\boldsymbol{n}\|_0 \leq \epsilon_0 \end{aligned} \quad (8)$$

To modify only one pixel, ϵ_0 is set to 1, hence, making the optimization problem hard. So, the authors applied the concept of Differential Evolution [65] to find the optimal solution. This technique requires the probabilistic labels predicted by the targeted model and does not necessitate any information about the network parameter values or gradients. It is implemented in a simple evolutionary strategy yet successfully fooling networks.

6) DEEPFOOL

Moosavi-Dezfooli *et al.* [66] proposed an iterative manner, called *DeepFool*, to find a minimal norm adversarial perturbation for a clean input image. The proposed algorithm initializes with the assumption that the input image is located in a region confined by the decision boundaries of an affine classifier, and the class label of the input is initially decided. At each iteration, the image is perturbed by a small vector. It is sought to lead the resulting perturbed image to the boundaries obtained by linearly approximating the region boundaries within which the image resides. In each iteration, the perturbations are added to the image and accumulated to compute the ultimate perturbation, which alters the input image label according to the original decision boundaries of the image region. *DeepFool* has been demonstrated to provide smaller perturbations compared to *FGSM* and *JSMA* while having similar fooling ratios.

7) UNIVERSAL ADVERSARIAL PERTURBATIONS

In contrast to their *DeepFool* method that computes image-specific perturbations, Moosavi-Dezfooli *et al.* [67] proposed their newer algorithm to generate image-agnostic *Universal Adversarial Perturbations* to fool a network on any image successfully. They attempted to find a universal perturbation that satisfies the following constraint:

$$\begin{aligned} &P\big(\boldsymbol{F}(\boldsymbol{x}) \neq \boldsymbol{F}(\boldsymbol{x} + \boldsymbol{n})\big) \geq \delta \\ &\text{s.t.}\ \ \|\boldsymbol{n}\|_{\mathrm{p}} \leq \xi \end{aligned} \quad (9)$$

where $P(.)$ denotes the probability, δ controls the fooling rate, $\|.\|_p$ refers to L_p-norm, and ξ confines the size of universal perturbation. Accordingly, the smaller the value of ξ, the more imperceptible the adversarial example to human eyes. It is shown that the *Universal Adversarial Perturbations* could be generalized well across popular deep learning architectures (e.g., VGG, CaffeNet, GoogLeNet, ResNet).

8) CARLINI & WAGNER ATTACKS (C&W)

Carlini and Wagner [68] introduced a set of adversarial attacks to defeat defensive distillation. According to their study, the L_0-, L_1- and L_2-norms of quasi-imperceptible perturbations are restricted to fail defensive distillation for the targeted networks. It is also demonstrated that the adversarial examples generated with un-distilled networks transfer well to the distilled networks making the generated perturbations proper for black-box attacks. Regarding definition, distillation is referred to as a training procedure to transfer knowledge of a more complex network to a smaller network. This notion was initially introduced by Hinton *et al.* [69]. Later, Papernot *et al.* [70] introduced the variant of the procedure using the knowledge of the network to improve its robustness.

IV. ADVERSARIAL EXAMPLE GENERATION AGAINST FACE RECOGNITION

In this section, we review adversarial examples generated against FR systems. We first explain the main attack generation methods introduced in the literature. Next, we compare different attacks according to their orientation. Finally, we repeat the comparison this time based on attributes of the adversarial capacity, specificity, transferability, and the perturbation type.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

A. METHODS

In this section, we review the main adversarial example generation methods against FR models. We review different studies in which they will be compared in succeeding sections to maintain the discussion flow.

1) IMAGE-LEVEL GRID-BASED OCCLUSION

Distortions that are not specific to faces and can be applied to any object image are categorized as image-level distortions. Goswami *et al.* [71] introduced an image-level distortion called *Grid-based Occlusion*. In this approach, points $\boldsymbol{P} = \{p_1, p_2, \ldots, p_n\}$ are selected along the image upper ($y = 0$) and left ($x = 0$) boundaries according to a parameter ρ_{grids}, where grids refer to *Grid-based Occlusion*. The ρ_{grids} parameter determines the number of grids utilized to alter the given image with higher values to result in a denser grid, i.e., more grid lines. For each point $p_i = (x_i, y_i)$, a point on the opposite boundary of the image, $p_i^{'} = (x_i^{'}, y_i^{'})$, is selected, with the condition if $y_i = 0$ then $y_i^{'} = H$, and if $x_i = 0$ then $x_i^{'} = W$, where $W \times H$ is the input image size. Once a set of pair points $\boldsymbol{P}$ and $\boldsymbol{P}^{'}$ selected, one-pixel wide lines are created to link each pair. Finally, the pixels placed on these lines set to 0 grayscale value.

2) IMAGE-LEVEL MOST SIGNIFICANT BIT-BASED NOISE (XMSB) DISTORTION

Image-level most significant bit-based noise is another image-level distortion introduced by Goswami *et al.* [71]. In this approach, three sets of pixels $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{X}_3$ are selected stochastically from the image such that $|\mathcal{X}_i| = \emptyset_i \times W \times H$. Here $W \times H$ is the input image size, and the parameter $\emptyset_i$ represents the fraction of pixels where the i^{th} most significant bit is flipped. Accordingly, the higher the value of $\emptyset_i$, the more pixels are distorted in the i^{th} most significant bit. For each $\mathcal{P}_j \in \mathcal{X}_i$, $\forall i \in [1,3]$, the following operation is pursued:

$$\mathcal{P}_{\mathrm{kj}} = \mathcal{P}_{\mathrm{kj}} \oplus 1 \tag{10}$$

where $\mathcal{P}_{kj}$ represents the k^{th} most significant bit of the j^{th} pixel in the set and $\oplus$ denotes the bitwise XOR operation. Also, it should be noted that the sets $\mathcal{X}_i$ may overlap; hence, the total number of pixels influenced by the noise is less than or equal to $|\mathcal{X}_1| + |\mathcal{X}_2| + |\mathcal{X}_3|$, depending on the stochastic selection.

3) FACE-LEVEL DISTORTION

Besides image-level distortion, Goswami *et al.* [71] also introduced face-level distortions. This type of distortion expressly necessitates face-specific information, e.g., location of facial landmarks. As a result, this approach is typically applied after performing automatic face and facial landmark detection. Once facial landmarks are detected, they are utilized along with their boundaries to perform the masking step. To obscure the eye region, a singular blocking band is drawn on the face image as follows:

$$I\{x, y\} = 0,\ \forall x \in [0, W], y \in \left[y_e - d_{eye}/\psi\ ,\ y_e + d_{eye}/\psi\right] \tag{11}$$

where $y_e = (y_{le} + y_{re})/2$, and (x_{le}, y_{le}) and (x_{re}, y_{re}) are positions of left eye center and right eye center, respectively. The d_{eye} is the inter-eye distance and calculated as $x_{re} - x_{le}$, and ψ is the parameter that determines the occlusion band's width. The *Eye Region Occlusion* (*ERO*) process could be implemented to obscure the forehead and brow in a similar trend using the facial landmarks on the forehead and brow regions as a mask. It could also be implemented to occlude the beard region utilizing the outer facial landmarks and nose and mouth coordinates to create the mask as combinations of individually occluded areas.

4) EVOLUTIONARY ATTACK

Dong *et al.* [72] proposed *Evolutionary Attack* method, based on (1+1)-CMA-ES [73], which is a useful and straightforward variant of the covariance matrix adaptation evolution strategy (CMA-ES) [74]. In each update iteration of the (1+1)-CMA-ES, a new offspring (candidate solution) is generated from its parent (current solution) by adding random noise, the objective of these two solutions is evaluated, and the better one is selected for the next iteration. This method can solve the black-box optimization problem of:

$$\min_{\boldsymbol{x}'} L(\boldsymbol{x}') = \left\|\boldsymbol{x}' - \boldsymbol{x}\right\|_2 + \delta\left(\mathcal{C}\left(\boldsymbol{F}(\boldsymbol{x}')\right) = 1\right) \tag{12}$$

where $\mathcal{C}(.)$ is an adversarial criterion that takes 1 if the attack requirement is satisfied and 0 otherwise, and $\delta(a)$ is 0 if a is true, and $+\infty$, otherwise. However, the authors did not apply the (1+1)-CMA-ES to optimize (12) due to the high dimension of $\boldsymbol{x}^{'}$. To accelerate this algorithm, they proposed an appropriate distribution to sample the random noise in each iteration, which can model the local geometry of the search directions. They sampled a random noise from a biased Gaussian distribution to minimize the distance of the sampled adversarial image from the original image. This added bias term is a critical hyper-parameter controlling the strength of going towards the original image. The authors also proposed techniques to reduce the dimension of search space by considering the characteristics of this problem. They sampled random noise in a lower-dimensional space $\mathbb{R}^m$ with $m < d$, where d is the dimension of input space. They then adopted an upscaling operator, precisely, the bilinear interpolation method, to project noise vector to the original space. Consequently, the input image dimension is preserved, and the dimension of search space is reduced.

5) FEATURE FAST & ITERATIVE ATTACK METHODS

Given a face pair and a deep face model, [75] proposed feature-level attacks to compare the face pair via calculating the distance between their normalized deep representations. These representations are similar to the embedding features, except that they are normalized and extracted from the deep face model. To discover the vulnerability of deep face models, the authors proposed to add perturbation on one of the face images to generate adversarial examples and deceive the face model. According to their notion, a positive and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

negative face pair is defined, for which the corresponding output labels are the same and different, respectively. Denoting the face pair by $\{\boldsymbol{x^1},\ \boldsymbol{x^2}\}$ and adversarial example by $\boldsymbol{x}' = \boldsymbol{x}^1 + \boldsymbol{n}$, for a positive face pair, $l^1 = l^2$ and the optimized objective and loss function are formulated as:

$$\begin{aligned} \boldsymbol{n} &= \underset{\boldsymbol{n}}{argmax}\ \|\boldsymbol{F}(\boldsymbol{x}^1 + \boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{x^2})\|_2,\ \|\boldsymbol{n}\|_\infty < \varepsilon \\ \mathcal{J}(\boldsymbol{x}^1 + \boldsymbol{n}, \boldsymbol{x}^2) &= \ \|\boldsymbol{F}(\boldsymbol{x}^1 + \boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{x}^2)\|_2 \end{aligned} \tag{13}$$

whereas for negative face pair $\{\boldsymbol{x}^1, \boldsymbol{x}^2\}$, $l^1 \neq l^2$ and the optimized objective and loss function are formulated as:

$$\begin{aligned} \boldsymbol{n} &= \underset{\boldsymbol{n}}{argmax}\ \|\boldsymbol{F}(\boldsymbol{x}^1 + \boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{x}^2)\|_2,\ \|\boldsymbol{n}\|_\infty < \varepsilon \\ \mathcal{J}(\boldsymbol{x}^1 + \boldsymbol{n}, \boldsymbol{x}^2) &= \ -\|\boldsymbol{F}(\boldsymbol{x}^1 + \boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{x}^2)\|_2 \end{aligned} \tag{14}$$

where $F(\boldsymbol{x}^i)$ denotes deep representations after normalization and ε limits the maximum deviation of the perturbation. Forming adversarial perturbation based on the loss functions of (13) and (14) is called *Feature Fast Attack Method (FFM)* and defined as:

$$\boldsymbol{x}^1 + \boldsymbol{n} = \mathcal{G}_{\boldsymbol{x}^1, \varepsilon}\left(\boldsymbol{x}^1 + sign\left(\nabla_{\boldsymbol{x}^1}\mathcal{J}(\boldsymbol{x}^1, \boldsymbol{x}^2)\right)\right) \tag{15}$$

Considering an iterative way, the authors proposed the *Feature Iterative Attack Method (FIM)* as:

$$\begin{aligned} &\boldsymbol{n_0} = 0 \\ &\mathrm{g}_{\mathrm{N}+1} = \nabla_{\mathbf{x}^1+\mathbf{n}_\mathrm{N}}\mathcal{J}(\boldsymbol{x}^1 + \boldsymbol{n}_N, \boldsymbol{x}^2) \\ &\boldsymbol{x}^1 + \boldsymbol{n}_{N+1} = \left(\boldsymbol{x}^1 + \boldsymbol{n}_N, \mathrm{sign}(\mathrm{g}_{\mathrm{N}+1})\right) \end{aligned} \tag{16}$$

where $\mathcal{G}_{\boldsymbol{x},\varepsilon}(\boldsymbol{x}') = min(255,\ \boldsymbol{x} + \varepsilon, max(0,\ \boldsymbol{x} - \varepsilon,\ \boldsymbol{x}'))$; the iteration can be chosen heuristically $min(\varepsilon + 4,\ 1.25\varepsilon)$.

6) EYEGLASS ACCESSORY PRINTING

Sharif *et al.* [76] proposed a physically realizable attack for impersonation or dodging in a digital environment. To enable physical realizability, the first step involved implementing the attacks purely with facial accessories (specifically, eyeglass frames) via 3d- or even 2d-printing technologies. In particular, they used a specific readily available digital model of eyeglass frames and utilized a commodity inkjet printer (Epson XP-830) to print the front plane of the eyeglass frames on glossy paper, which are affixed to actual eyeglass frames, subsequently. After alignment, the frames occupy about 6.5% of the 224 × 224 face image pixels, implying that the attacks perturb at most 6.5% of the pixels in the image. To find the color of the frames necessary to achieve impersonation or dodging, their color is initialized to a solid color (e.g., yellow), and the frames are rendered onto the image of the subject. Their color is updated iteratively through the gradient descent process to craft adversarial perturbations tolerant to slight natural movements when physically wearing the frames.

The second step involved tweaking the mathematical formulation of the attacker's objective to focus on adversarial perturbations that both robust to small changes in viewing condition and smooth as expected from natural images. To find perturbations independent of the exact imaging conditions, aiming to enhance the generality of the perturbations, the authors looked for perturbations that can cause any image in a set of inputs to be misclassified. To this end, an attacker collects a set of images, $\boldsymbol{X}$, and finds a single perturbation that optimizes her objective for every image $\boldsymbol{x} \in \boldsymbol{X}$. For impersonation, this is formalized as the following optimization problem (dodging is analogous):

$$\underset{\boldsymbol{n}}{argmin} \sum_{x \in X} softmaxloss(\boldsymbol{F}(\boldsymbol{x} + \boldsymbol{n}),\ l) \tag{17}$$

where $\boldsymbol{n}$ denotes the perturbation. To preserve the smoothness of perturbations, the optimization is updated to account for minimizing total variation (TV) [77], which is defined as:

$$\begin{aligned} TV(\boldsymbol{n}) = \sum_{i,j} \Big(&\left(\boldsymbol{n}_{i,j} - \boldsymbol{n}_{i+1,j}\right)^2 \\ &+ \left(\boldsymbol{n}_{i,j} - \boldsymbol{n}_{i,j+1}\right)^2 \Big)^{1/2} \end{aligned} \tag{18}$$

where $\boldsymbol{n}_{i,i}$ denotes a pixel in $\boldsymbol{n}$ at coordinate (i, j). $TV(\boldsymbol{n})$ is low when the values of adjacent pixels are close to each other (i.e., the perturbation is smooth), and high otherwise. Therefore, by minimizing $TV(\boldsymbol{n})$, the smoothness of the perturbed image hence the physical realizability is improved.

7) VISIBLE LIGHT-BASED ATTACK (VLA)

Shen *et al.* [78] introduced a *Visible Light-based Attack* (*VLA*) against FR systems, where visible light-based adversarial perturbations are crafted and projected on human faces. For each adversarial example, the authors proposed to generate a perturbation frame and a concealing frame, which are projected to the face of the user. The perturbation frame contains information on how to change the input user's facial features to the features of a targeted or non-targeted user, whereas the concealing frame aims to hide the perturbations in the perturbation frame from being observed by human eyes.

Regarding the perturbation frames generation, this method enlarges the pixel-level image modifications into region-level to avoid probable perturbation loss in physical scenarios. Accordingly, the perturbation frame is divided into exclusive ranges based on the similarity of containing color values. A Manshift clustering divides all colors, where nearby similar colors are divided into the same regions, and each group of nearby pixels with the same color in the image is regarded as one perturbation region. Then, in the second step, a region filtering strategy is utilized to ensure that the camera can successfully capture all projected details in a perturbation frame, and small color regions would not get lost in the images captured in physical scenarios. Denoting $\boldsymbol{n} = \boldsymbol{x}' - \boldsymbol{x}$ as the perturbation frame, a clustering and filtering result of $\boldsymbol{n}$ is denoted by $\boldsymbol{C}_{\boldsymbol{x},\boldsymbol{x}'}$ and defined as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

$$\boldsymbol{C}_{\boldsymbol{x},\boldsymbol{x}'} = \{G_i(p),\ R_i | 0 \le i \le m\} \tag{19}$$

where $G_i(p)$ indicates whether the color of a pixel p should be set as R_i, and m is the total number of color regions. For each pixel $\boldsymbol{p}$. in the image $\boldsymbol{C}_{\boldsymbol{x},\boldsymbol{x}'}$, $G_i(p)$ is 1 if p lies within R_i, and 0, otherwise. The generation function $H(\cdot)$ is defined next to transform the clustering result $\boldsymbol{C}_{\boldsymbol{x},\boldsymbol{x}'}$ into a perturbation frame $\boldsymbol{n}$, as shown in (20):

$$\boldsymbol{n} = H\left(\boldsymbol{C}_{\boldsymbol{x},\boldsymbol{x}'}\right) = [R_i \text{ if } G_i(p) = 1] \tag{20}$$

To hide the perturbation frames from human eyes, concealing frames are generated according to the effect of persistence of vision (POV) [79]. According to POV, two different colors that swap frequently cause the human brain not directly process these changes at the exact moment they occur, making the human eyes perceive a new color as a fusion of those colors. Based on this knowledge, by projecting the perturbation frame and the concealing frame alternately, i.e., displaying the corresponding two colors of generated images interchangeably, it can be difficult for human eyes to feel the perturbation frame, and a fusion of these colors will be perceived as a base/background color of the image.

8) ADVHAT ATTACK

Komkov and Petiushko [80] proposed a reproducible adversarial attack generation method, called *AdvHat*. They printed a rectangular paper sticker on a standard color printer and put it on the hat with an off-plane transformations algorithm. The proposed algorithm split into two steps: (1) off-plane bending of the sticker, which is simulated as a parabolic transformation in the 3D space to map each point of the sticker to the new point on the parabolic cylinder, and (2) pitch rotation of the sticker, which is stimulated by the application of a 3D affine transformation to the obtained new points. The authors projected the resulted sticker on the high-quality face image with small perturbations in the projection parameters. They transformed the new face image into the standard template of ArcFace input to pass it to the optimization step. Regarding the optimization step, the sum of two parameters (TV loss and cosine similarity between two embeddings) is minimized as follows to achieve the gradient signs used to modify the sticker image:

$$L_T(\boldsymbol{x}', \boldsymbol{a}) = L_{\text{sim}}(\boldsymbol{x}', \boldsymbol{a}) + \lambda \cdot TV(patch) \tag{21}$$

where L_T is the total loss, $patch$ denotes the sticker, $\boldsymbol{x}'$ is a photo with the applied patch, and λ is a weight for TV loss, which is assumed to be $1e-4$ in this work. Here, L_{sim} is cosine similarity between two embeddings and defined as follows:

$$L_{\text{sim}}(\boldsymbol{x}', \boldsymbol{a}) = \cos\left(e_{\boldsymbol{x}'}, e_{\boldsymbol{a}}\right) \tag{22}$$

where $e_{\boldsymbol{x}'}$ is obtained embeddings of the face image of the attacker and e_a refers to the embedding of the desired person's face image calculated by ArcFace.

9) PENALIZED FAST GRADIENT VALUE METHOD (P-FGVM)

Chatzikyriakidis *et al.* [81] introduced a *Penalized Fast Gradient Value Method* (*P-FGVM*) adversarial attack technique, which runs on the image spatial domain and generates adversarial de-identified facial images like the original ones. This technique is inspired by the *I-FGVM*, with a minor exception of combining an adversarial loss and a "realism" loss term in its gradient descent update equations. In this method, a targeted adversarial example $\boldsymbol{x}'$ is generated through the following gradient descent update equations:

$$\begin{aligned}\boldsymbol{x}'^{(i+1)} = Clip_{\epsilon}\Big\{\boldsymbol{x}'^{(i)} + \alpha \\ \cdot \left(\nabla_{\boldsymbol{x}'^{(i)}} \mathcal{J}\left(\boldsymbol{\theta}, \boldsymbol{x}'^{(i)}, l\right)\right. \\ \left. + \lambda\left(\boldsymbol{x}'^{(i)} - \boldsymbol{x}\right)\right)\Big\}\end{aligned} \tag{23}$$

where λ is a weight coefficient and $(\boldsymbol{x}'^{(i)} - \boldsymbol{x})$ is the realism loss term.

10) FACE FRIEND-SAFE ATTACK

Kwon *et al.* [82] proposed the *Face Friend-safe* adversarial example generation method, which generates adversarial examples that are misrecognized by an enemy FR system, nonetheless, appropriately recognized by a friend FR system with the least distortion. The proposed method consists of a transformer, a friend classifier M_{friend}, and an enemy classifier M_{enemy}, to generate adversarial face images. Given the pre-trained M_{friend} and M_{enemy} and the original input $\boldsymbol{x} \in \boldsymbol{X}$, the optimization problem of generating the adversarial face example $\boldsymbol{x}'$ is as follows:

$$\begin{aligned}&\underset{\boldsymbol{x}'}{argmin}\ L(\boldsymbol{x}, \boldsymbol{x}') \\ &s.t.\ \ \mathrm{g}^{\text{friend}}(\mathrm{x}') = 1 \ \text{ and } \ g^{\text{enemy}}(\mathrm{x}') \neq 1\end{aligned} \tag{24}$$

where $g^{\text{friend}}(\boldsymbol{x})$ and $g^{\text{enemy}}(\boldsymbol{x})$ denote the operation functions of a friend classifier M_{friend} and enemy classifier M_{enemy}, respectively. $L(.)$ is the distance measured between the face original sample $\boldsymbol{x}$ and face transformed example $\boldsymbol{x}'$. The transformer generates adversarial face example $\boldsymbol{x}'$, taking the original sample $\boldsymbol{x}$ and its corresponding output label. The classification loss of $\boldsymbol{x}'$ by M_{friend} and M_{enemy} are returned to the transformer, which then calculates the total loss, L_T, and repeats the above procedure to generate an adversarial face example $\boldsymbol{x}'$ while minimizing L_T. This total loss is defined as follows:

$$L_T = L_{\text{friend}} + L_{\text{enemy}} + L_{\text{distortion}} \tag{25}$$

where L_{friend} is the classification loss function of M_{friend}, L_{enemy} is the classification loss function of M_{enemy}, and $L_{\text{distortion}}$ is the distortion of the transformed example, and defined as the distance between $\boldsymbol{x}$ and $\boldsymbol{x}'$.

11) FAST LANDMARK MANIPULATION (FLM) METHOD

Dabouei *et al.* [83] proposed a fast landmark manipulation approach to craft adversarial faces. They proposed to generate adversarial examples by spatially transforming

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

original images. Using a landmark detector function $\boldsymbol{\Phi}$, that maps the face image $\boldsymbol{x}$ to a set of k 2D-landmark locations $\boldsymbol{P} = \{p_1, \dots, p_k\}$, $p_i = (u_i, v_i)$, it is assumed that $p_i^{'} = (u_i^{'}, v_i^{'})$ is the transformed version of p_i, and defines the i^{th} landmark location in the corresponding adversarial image $\boldsymbol{x}'$. To manipulate the face image based on $\boldsymbol{P}$, a per-landmark flow (displacement) f is defined to produce the location of the corresponding adversarial landmarks. Accordingly, the adversarial landmark $p_i^{'}$ can be obtained from the original landmark p_i and optimized particular displacement vector $f_i = (\Delta u_i, \Delta v_i)$ as follows:

$$\begin{aligned} \mathrm{p}_\mathrm{i}^{'} &= \mathrm{p}_\mathrm{i} + \mathrm{f}_\mathrm{i} \\ (u_i^{'}, v_i^{'}) &= (u_i + \Delta u_i,\ v_i + \Delta v_i) \end{aligned} \tag{26}$$

In contrast with the reference work [84], which fulfills this purpose by defining field f for all pixel locations in the input image, Dabouei *et al.* [83] defined it only for k landmarks, which is notably small compared to the number of pixels in the input image, especially when incorporated in real applications like FR problems. This limited number of control points also reduces the distortion introduced by the spatial transformation. Using the transformation $\boldsymbol{T}$, the benign face image spatially transformed into an adversarial face image as follows:

$$\boldsymbol{x}' = T(\boldsymbol{P},\ \boldsymbol{P}', \boldsymbol{x}) \tag{27}$$

where $\boldsymbol{P}'$ refers to target control points. Incorporating the softmax cost as the measure for the correct classification, authors defined the total loss for generating adversarial faces as:

$$\begin{aligned} L(\boldsymbol{P}, \boldsymbol{P}',\ \boldsymbol{x},\ l) = softmaxloss\big(\boldsymbol{F}\big(T(\boldsymbol{P},\ \boldsymbol{P}', \boldsymbol{x})\big),\ l\big) \\ - \lambda_{flow} L_{flow}(\boldsymbol{P}' - \boldsymbol{P}) \end{aligned} \tag{28}$$

where λ_{flow} is a positive coefficient used to control the magnitude of displacement, and L_{flow} is a term incorporated for bounding the displacement field. This way, the landmark displacement field f is found iteratively using the gradient direction of the prediction and called the *FLM* method. Authors also extended this approach proposing the *Grouped Fast Landmark Manipulation* (*GFLM*) Method, which semantically groups landmarks and manipulates the group properties instead of perturbing each landmark. This idea was formed to resolve severe distortion of the adversarial faces generated by *FLM* and preserve the whole structure of the created images.

B. COMPARISON OF DIFFERENT ADVERSARIES ON ORIENTATION

A general taxonomy of existing adversarial example generation techniques against FR systems considering the orientation of adversaries is depicted in Fig. 3. Based on the strategies followed in different studies or tools recruited to launch adversarial attacks, different techniques could be mainly classified into four categories, namely, (1) CNN models-oriented; (2) physical attacks-oriented; and (3) geometry-oriented. The remainder of this section is structured according to this classification.

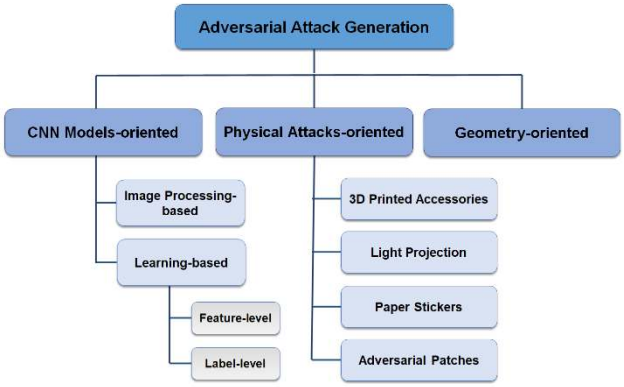

FIGURE 3. The broad categorization of adversarial attack generation methods aimed to deceive the FR systems.

1) CNN MODELS-ORIENTED

As stated earlier, deep learning paradigm has seen a remarkable propagation in the FR mission. Several models are deep CNN-based architectures with many hidden layers and millions of parameters, which are designed to achieve very high accuracies when tested on different databases. Whereas reported efficiencies of such models improve progressively, they are shown to be susceptible to adversarial attacks. Realizing this, many researchers have started to design approaches to exploit the weaknesses of such algorithms.

Goswami *et al.* [71] considered the vulnerability of several deep CNN-based FR algorithms in the presence of image processing-based distortions at (1) image-level and (2) face-level. They confirmed that attacks on systems do not need to be sophisticated learning based. Instead, a random noise or even horizontal and vertical black grid lines drawn in the face image can severely reduce the face verification accuracies. Examples of this effort are depicted in Fig. 4.

FIGURE 4. Clean input images (a) modified by image processing-based distortions of *xMSB* (b), *Grid-based Occlusion* (c), *Forehead and Brow Occlusion* (*FHBO*) (d), *Eye Region Occlusion* (*ERO*) (e), and *Bread-like Occlusion* (f) [71].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

Dong *et al.* [72] proposed the *Evolutionary Attack* algorithm to evaluate the robustness of multiple advanced FR models against label-level adversarial examples in a decision-based attack setting.

Zhong and Deng [85] defined *Dropout Face Attacking Networks* (*DFANet)* technique to explore the vulnerability of deep CNNs against feature-level adversarial examples. They incorporated dropout in the convolutional layers in the iterative steps of the adversarial generation process to improve the transferability of adversarial examples. Specifically, for a face model composed of convolutional layers, given the output of the i^{th} convolutional layer, they proposed to generate a mask with elements that independently sampled from a Bernoulli distribution. This mask is then utilized to modify the output of the i^{th} convolutional layer via Hadamard product of those. Authors proposed to apply this method to the generation of *FIM* and combined it with transferability enhancement methods [86]–[88]. Applying their practice on the LFW dataset, they generated a new set of adversarial face pairs to attack commercial APIs of Amazon[3], Microsoft[4], Baidu[5], and Face++[6], which provide highly accurate facial analysis and facial search capabilities to detect, analyze, and compare faces for a wide variety of applications. They made this TALFW database available to the public for future investigations.

Garofalo *et al.* [89] focused on the security aspect of face authentication systems aiming to let impostors evade the FR models. The authors deployed a poisoning attack on an authenticator based on the OpenFace FR framework which was extended with a support vector machine (SVM) classifier. They implemented the attack against the underlying SVM model to classify face templates extracted by the FaceNet model. In another study with a similar purpose, Chatzikyriakidis *et al.* [81] proposed to utilize adversarial examples in cases of face de-identification. They introduced the *P-FGVM* adversarial attack technique against CNN-based face classifiers. Examples of implementing this method to generate adversarial images are shown in Fig. 5.

Lately, Kwon *et al.* [82] proposed the *Face Friend-safe* adversarial example generation method to successfully mislead an enemy FR system, nonetheless, be appropriately recognized by a friend FR system.

Recently, a new Python-based toolbox, termed Advbox, is proposed to generate adversarial examples [90]. With Advbox, it is possible to fool neural networks in PaddlePaddle, PyTorch, Caffe2, MxNet, Keras, and TensorFlow, with the additional capability to benchmark the robustness of ML models. Compared to previous works, this platform supports actual attack scenarios, such as FR attacks.

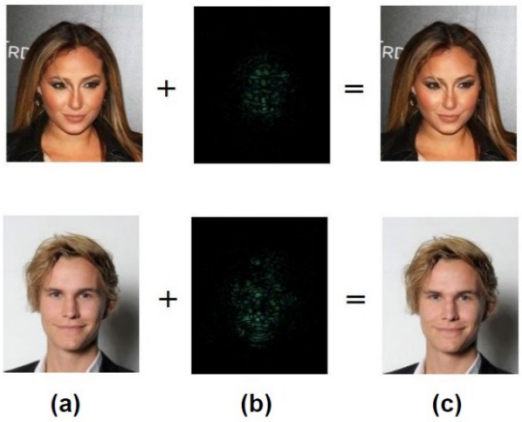

FIGURE 5. Clean facial images (a) modified by adversarial perturbation (b) to generate de-identified facial images (c) via adversarial attack method *P-FGVM* [81]. The absolute value of perturbation is amplified by 10x.

2) PHYSICAL ATTACKS-ORIENTED

Intruders to facial biometric systems often encountered two kinds of challenges: (1) they do not have precise control over the FR systems' (digital) input; instead, they may be able to control their physical appearance, and (2) they might be easily observed by traditional means like the police, when manipulating their appearances to evade recognition, e.g., with an excessive amount of makeup. In the light of such challenges, a new class of adversarial attacks has emerged based on the physical state of the attackers.

Sharif *et al.* [76] developed the *Eyeglass Accessory Printing* method to generate a physically realizable yet inconspicuous class of attacks. In [91], authors proposed *Adversarial Generative Nets* (*AGNs*) to generate images of artifacts (e.g., eyeglasses) that would lead to misclassification. The artifacts generated by such neural networks resembled a reference set of artifacts (e.g., real eyeglass designs) and satisfied the inconspicuousness objective. Similar to GANs, *AGNs* are adversarially trained against a discriminator to learn how to generate realistic images. Differently from GANs, *AGNs* are also trained to generate adversarial outputs that can mislead given FR models on both digital and physical levels of evasion purposes. In this study, the FR algorithms were targeted on the digital-level by traditional attacks, such as Szegedy's *L-BFGS* method [9], and deceived on the physical-level by requesting individuals to wear their 3D-printed sunglasses frames. Fig. 6 illustrates an impersonation attack generation by wearing such an accessory.

Zhou *et al.* [92] designed a cap, with some penny-size lit Infrared LEDs on the peak, to generate inconspicuous physical adversarial attacks via Infrared dot direction on the carrier's face. The loss in this work is optimized by adjusting light spots in line with the model on the attacker's photo. The

[3] https://aws.amazon.com/rekognition
[4] https://azure.microsoft.com
[5] https://ai.baidu.com
[6] https://www.faceplusplus.com.cn

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

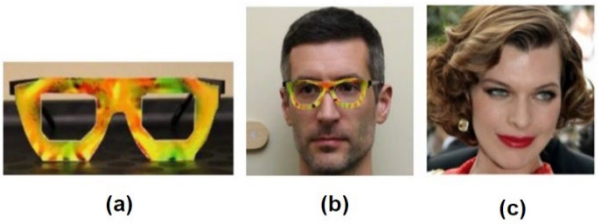

FIGURE 6. **The eyeglass frames (a) were used by Lujo Bauer (b) to impersonate Milla Jovovich (c) [76].**

attacker could then evade detection by adjusting the positions, sizes, and strengths of the dots.

Motivated by the differences in image-forming principles between cameras and human eyes, Shen *et al.* [78] proposed the *VLA* attack against FR models. In a similar study, Nguyen *et al.* [55] studied the feasibility of directing real-time physical attacks on FR systems by adversarial light projections using a web camera and a projector. In this approach, the authors captured the adversary's facial image with a camera and used one or more target images to (1) adjust the camera-projector setup according to the attack environment and (2) create a digital adversarial pattern. The digital pattern is then projected onto the adversary's face in the physical domain with a projector to evade recognition. Although this work's objectives are identical to the infrared-based adversarial attacks [92], it does not necessitate creating a wearable artifact, thus, offers a more comfortable alternative setup to direct physical attacks on FR models.

Another study [80] proposed to target the public Face ID model LResNet100E-IR, ArcFace@ms1m-refine-v2, by *AdvHat* attack generation method in fixed (full-face photos with uniform light) and variable (different angles of the face rotation and light conditions) settings. Similarly, Pautov *et al.* [93] examined the security of the same recognition system and proposed to print, add (as face attributes) and photograph adversarial patches; the snapshot of an individual with such attributes is then delivered to the classifier to alter the correctly recognized class to the desired one. In this work, patches were either various parts of the attacker's face, like nose or forehead or some wearable accessories such as eyeglasses.

3) GEOMETRY-ORIENTED

Prevalent intensity-based adversarial attack methods, which manipulate the intensity of input images directly, are computationally cheap but sensitive to spatial transformations. A small rotation, translation, or scale variation in the input image could result in a drastic change in similarity in these methods. Due to this limitation, a new class of attacks was initiated to generate geometry-based adversarial examples.

Dabouei *et al.* [83] proposed the *FLM* method to craft adversarial faces almost 200 times quicker than traditional geometric attacks. They further introduced *GFLM* as the extended version of the fast geometric perturbation generation algorithm. Fig. 7 demonstrates an overview of the proposed fast geometry-based adversarial attack [83].

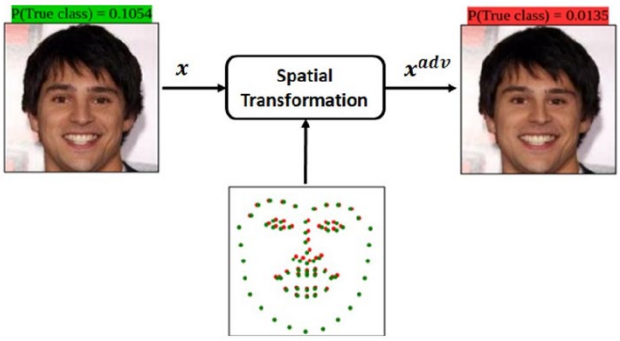

FIGURE 7. **Fast landmark manipulation method application to produce adversarial landmark locations, with which the ground truth image spatially transformed to a natural adversarial image. As shown in green and red colors, the ground truth image is correctly classified, whereas the adversarial image is misclassified to a wrong class [83].**

Song *et al.* [94] focused on attacks that mislead the FR networks to detect someone as a target person, not misclassify inconspicuously. They introduced an *Attentional Adversarial Attack Generative Network* (A^3GN) to generate adversarial examples similar to the original images while having the same feature representation as to the target face. To capture the target person's semantic information, they appended a conditional variational autoencoder and attention modules to learn the instance-level correspondences between faces.

Utilizing GANs, Deb *et al.* [56] crafted natural face images with a barely distinguishable difference from target face images. They proposed the *AdvFaces* adversarial face synthesis method to craft minimal perturbations in the prominent facial regions. This method comprises a generator, a discriminator, and a face matcher to automatically generate an adversarial mask added to the image to obtain an adversarial face image. Table I. presents a general overview of different adversarial example generation approaches regarding their orientation.

C. COMPARISON OF DIFFERENT ADVERSARIES ON EVALUATION PROCESS

This section compares different adversarial example generation techniques in terms of their evaluation process and corresponding utilized metrics.

Goswami *et al.* [71] evaluated the verification performance of CNN-based FR algorithms, including OpenFace, VGG-Face, LightCNN, and L-CSSE [95], and one commercial-off-the-shelf recognizer (COTS) in the presence of image processing based adversarial distortions on the PaSC [96] and MEDS [97] databases. They reported experimental results based on the genuine accept rate (GAR) (%) of the attacks at 1% false accept rate (FAR). Overall, they demonstrated that deep learning-based algorithms could experience higher performance drop as opposed to the non-deep learning-based COTS when any distortion is introduced in the data.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

TABLE I
COMPARISON OF DIFFERENT ADVERSARIAL ATTACK GENERATION ALGORITHMS ON THE ORIENTATION AND EVALUATION PROCESS

Representative study	Attack orientation	Method/Description	Evaluation metrics
[71]	CNN models	*(1) Image-level Grid-based Occlusion, (2) Image-level Most Significant Bit-based Noise (xMSB) Distortion, (3) Face-level Distortion*	GAR (%) @ 1% FAR
[72]	CNN models	*Evolutionary Attack*	MSE
[85]	CNN models	*DFANet*	Hit rate
[89]	CNN models	Poisoning attack on an authenticator based on OpenFace extended with an SVM classifier	FNR, FPR, CE
[81]	CNN models	*P-FGVM*	MSSIM
[82]	CNN models	*Face Friend-safe Attack*	SR of the enemy classifier, the accuracy of the friend classifier, and average distortion
[90]	CNN models	Advbox toolbox	
[76], [91]	Physical	*Eyeglass Accessory Printing*	SR
[92]	Physical	Physical adversarial example generation via an infrared LEDs-equipped cap	L_2 distance between feature vectors of given pair of faces
[78]	Physical	*VLA*	SR
[55]	Physical	Physical adversarial example generation via real-time light projection	Similarity score threshold @ 0.01% FAR
[80]	Physical	*AdvHat attack*	Baseline similarity and final similarity
[93]	Physical	Adversarial example generation by printing, adding, and photographing adversarial patches of nose, forehead, and eyeglasses of the attacker	Cosine similarity between embeddings of the given pair of faces
[83]	Geometric	*FLM*	SR, computation time
[94]	Geometric	Adversarial example generation via A^3GN	Physical likeness, similarity score, recognition accuracy
[56]	Geometric	Adversarial face generation via *AdvFaces* method	SR, SSIM

Dong *et al.* [72] compared the performance of the *Evolutionary Attack* method with all existing decision-based black-box attack generation methods, including the boundary attack method [98], optimization-based method [99], and an extension of NES in the label-only setting (NES-LO) [100]. On the LFW and MegaFace datasets, the authors made this comparison against SphereFace, CosFace, and ArcFace FR models. For all methods, they measured the distortion between the adversarial and original images by mean square error (MSE) to evaluate the performance of different methods. Experimental results demonstrated that the proposed method could converge much faster and achieve smaller distortions compared with other methods consistently across both tasks (*i.e.*, face verification and identification), both attack settings (*i.e.*, dodging and impersonation), and all face models.

Zhong and Deng [85] evaluated the transferability of targeted attacks between the ResNet-50 model trained on four datasets of CASIA-WebFace, MS-Celeb-1M, VGGFace2, and IMDb-Face [101]. They defined the goal of the attack as to generate adversarial examples from the source images and planned to obtain face embedding representations of source images closer to those of target images than the distance threshold of the FR systems. Accordingly, they computed the Euclidean distance of normalized deep features to obtain ROC curves and identified distance thresholds for judging whether a pair of source/target images is positive or negative. In this study, the attack is defined as a success (hit) when the embedding distance between the source image and target is less than the threshold. Authors used *Fast Target Gradient Sign Method* (*FTGSM*) [41] and *Iterative Target Gradient Sign Method* (*ITGSM*) [41] to generate label-level adversarial examples and *FFM* and *FIM* to generate feature-level adversarial examples. Being more effective in terms of the transferability, authors selected *FIM* as the baseline method and further improved it by incorporating the transferability enhancement methods [86]–[88]. Created strong baseline method was then compared with the proposed *DFANet* method. Based on the comparisons, the authors verified the superiority of the *DFANet* method and that most of the successful hit rates of adversarial examples generated by this approach could be improved to approximately 90% between the four deep FR models.

Garofalo *et al.* [89] utilized the Facescrub dataset [102] for their in-depth evaluation, as this dataset offers a high quantity of identities and samples per identity. They described the strength of the authenticator by false negative rate (FNR), false positive rate (FPR), and classification error (CE). Experimental results demonstrated that with the most successful attack, an impressive mean CE of 40.11% could be achieved, which was an increase in mean authentication error of almost 37% over the not targeted system, while the mean FPR increase was shown to be over 40%. In addition, the most successful attack deployment showed to lead to the CE of 51.23% on the test set, making the face authentication system entirely useless.

Chatzikyriakidis *et al.* [81] evaluated the proposed *P-FGVM* method on two CNN-based face classifiers: (1) a simple architecture model and (2) a fine-tuned model with transfer learning based on the pre-trained VGG-Face CNN

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

descriptor, using the VGG-16 architecture [43]. They calculated the mean structural similarity index (MSSIM) between the de-identified and original facial images as well as the L_2 norm of the adversarial perturbation as the metrics for measuring the visual quality of the results. Comparing with the baseline *I-FGVM* and *I-FGSM* methods, against the face classifiers described above and on a subset of the CelebA dataset, the authors demonstrated that the proposed method could produce de-identified images that are much closer to the original ones while having better misclassification error than the competing methods (3% and 1.7% increase in misclassification rate as compared with *I-FGVM* and *I-FGSM* methods, respectively).

Kwon *et al.* [82] considered the FaceNet recognition system as the target model; they trained their method on VGGFace2 and tested it on the LFW dataset. Authors evaluated the efficiency of the proposed method by measuring the attack success rate (SR) of the enemy classifier, the accuracy of the friend classifier, and the average distortion, demonstrating the values of 92.2%, 91.4%, and 64.22, respectively. Reporting such values, they claimed that the objectives of their work were achieved successfully.

Sharif *et al.* [76] evaluated their adversarial example generation method in both digital-environment and physical-realizability experiments. They measured the SR of the attack as the fraction of attempts to achieve the goal. To compute statistics that generalize beyond individual images, they performed each attack on three images of each subject and reported the mean SR across those images. In digital-environment experiments, attacking different DNNs under the white-box scenario, the attacker was able to dodge recognition or impersonate targets in almost all attempts with the mean SR of 100%. In Physical-realizability experiments, where subjects were asked to wear eyeglass frames and their images captured thereafter, the first three authors participated and for each of them, five sessions were considered. In the first session, the subjects did not wear the eyeglass frames, and non-adversarial images were classified correctly, with the mean probability of the correct class across the classification attempts above 0.85. In the second and third sessions, they wore eyeglass frames to attempt dodging against DNNs. The mean probability assigned to the subjects' class dropped remarkably from above 0.85 to less than 0.03, considering different cases. This was equivalent to achieving SRs of 100% (except for one experiment which resulted in an SR of 97.22%). In the fourth and fifth sessions, the subjects wore frames to attempt impersonation against DNNs. Considering different cases, more than 87.87% of the images collected in these sessions were misclassified by DNNs (with the mean probabilities of the targets greater than 0.75).

In [91], Sharif *et al.* assessed dodging and impersonation attacks against VGG-Face and OpenFace models. In the evaluation stage, they reported the accuracies of DNNs and SRs of the attacks. Using *AGNs,* in the digital domain all attempts succeeded with a mean SR of 100% in all dodging cases and greater than 88% in all impersonation attacks. In physical-realizability experiments, for dodging attacks, authors reported the *AGNs*' SR of 81% and 100% in the worst and best cases, respectively, and the mean probability assigned to the correct class of 0.40 and 0.01, correspondingly. For impersonation attacks, they reported the *AGNs'* SR of 53% and the mean probability assigned to the target of 0.22.

Zhou *et al.* [92] examined the effectiveness of their proposed technique against the FaceNet model on the LFW dataset. They used L_2 distance to weight the distance between two feature vectors generated by their model, and adopted the threshold 1.242 over the LFW dataset. In this way, a pair of faces with distance below the threshold were recognized as from the same person, otherwise two distinct individuals. The authors observed that the original distances, i.e., the distance between the embedding of the attacker and the victim before launching the attack, were all above the threshold. Hence, an authentication system could recognize that there was not a victim in the corresponding photo. On the other hand, the algorithm could result in adversarial examples that theoretically make distances fall below the threshold. In this work, theoretical distance means the distance between the calculated adversarial example and the victim. More importantly, the authors demonstrated that the attacker could indeed implement those adversarial examples by using the proposed device and consequently fool the authentication system. They verified this by measuring the distances after the attack that got below the threshold.

Shen *et al.* [78] conducted extensive experiments on the CusFace [78] and LFW datasets and against FaceNet, SphereFace, and dlib models. Authors generated adversarial examples using *FGSM* and *VLA* methods, separately. On the FaceNet model, they demonstrated that for the non-targeted attacks in physical scenarios, *VLA* could significantly improve the SR over the *FGSM*. For targeted attacks, however, the proposed method could achieve a reasonable SR. Experimental results explained that the region-level color areas in perturbation frames generated by the *VLA* are more robust helping to obtain more effective adversarial examples. Generated adversarial examples were also used to evaluate with other face recognizers of SphereFace and dlib. The results of *FGSM* indicated that the attack SR against dlib and SphereFace is less than that against FaceNet, as *FGSM* is a white-box approach and the adversarial examples targeting FaceNet may not fit for other recognizers. However, as the *VLA* is agnostic to face recognizers, it could exhibit a similar performance against the three recognizers.

Nguyen *et al.* [55] evaluated their approach against FaceNet, SphereFace, and one commercial-off-the-shelf FR system and confirmed the models' vulnerability to the light projection attacks. They used a similarity score threshold corresponding to FAR of 0.01% to determine if the attack is successful or not. Conducting impersonation and obfuscation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

experiments on live subjects and against the FaceNet system, the authors reported the highest SRs of 93.3% and 100%, respectively. This is while the lowest SR values were achieved against the commercial-off-the-shelf FR system, indicating the more vulnerability of the deep FR systems against generated attacks.

Komkov and Petiushko [80] evaluated the success and characteristics of the attacks in fixed and variable conditions. On the CASIA-WebFace dataset, they verified that their approach could easily confuse the LResNet100E-IR Face ID model. As the evaluation metrics, they explored baseline similarity and final similarity which they defined as cosine similarity between ground truth embedding and embedding for a photo with a hat, and cosine similarity between ground truth embedding and embedding for a photo with an adversarial sticker, respectively. In experiments with the fixed condition, they observed that adversarial stickers could significantly reduce the similarity to the ground truth class. In experiments with various conditions, where the robustness of the proposed approach to different shooting conditions aimed to be examined, although final similarity demonstrated to increase in each case/condition, the attack observed to work and almost all final similarities were shown to be less than the baseline similarities.

Pautov *et al.* [93] evaluated their method against ArcFace on CASIA-WebFace dataset and photos of the first and second authors of this work. They showed that with their simple attacking technique they could deceive the FR system in the digital and physical worlds. Experimental results demonstrated that though the similarity of the embedding corresponding to the photo of the attacker with an applied patch with ground truth class can reach just slightly below the similarity of that embedding with desired class, the FR model could not recognize the attacker as the ground truth class. Authors also discovered that the position of a patch, as well as its size, dramatically affects the success of the attack in the physical domain.

Dabouei *et al.* [83] evaluated the performance of the proposed *FLM* and *GFLM* methods for the white-box attack scenario. They trained the FaceNet model on VGGFace2 and CASIA-WebFace datasets and assessed its performance on the CASIA-WebFace dataset. The authors defined several experiments to investigate the importance of different regions of the face. From the results, they observed that with the attacks guided through these methods, the SR of more than 99.86% could be achieved. The computation time of these algorithms found to be noticeable too. The average time of generating adversarial faces for the *FLM* and *GFLM* was observed to be 125 and 254 milliseconds respectively, which is considerably shorter than the computation time of stAdv [84] (27.177 seconds on average).

Song *et al.* [94] examined the proposed method by training the model on CASIA-WebFace and evaluating it on LFW datasets. They compared their approach with stAdv and *GFLM* methods and observed that a satisfactory attack SR could be archived via their proposed method. Overall, the authors demonstrated the excellent performance of A^3GN by a set of evaluation criteria in physical likeness, similarity score, and accuracy of recognition on different target faces.

Deb *et al.* [56] quantified the effectiveness of their proposed adversarial example generation methods via attack SR and structural similarity index (SSIM). Authors trained *AdvFaces* on CASIA-WebFace and tested it on the LFW. They found that in comparison with the state-of-the-art adversarial example generation methods of *FGSM*, *PGD*, A^3GN, and *GFLM*, *AdvFaces* can generate adversarial faces similar to the test images to be matched against the gallery images. While evading the state-of-the-art FR models (FaceNet, SphereFace, ArcFace) and two commercial-off-the-shelf machers (COTS-A and COTS-B), generated images were demonstrated to attain attack SRs as high as 97.22% and 24.30% for obfuscation and impersonation attacks, respectively. They reported the structural similarities between adversarial and test images along with the time taken to generate a single adversarial image and demonstrated that with their proposed *AdvFaces* method, a computation time of 0.01 seconds and MSSIM of 0.95 and 0.92 could be achieved for obfuscation and impersonation attacks, respectively. Reported SSIM values and computation time were respectively higher and lower than those achieved by the other methods revealing the superiority of the *AdvFaces* method over them. Different evaluation metrics that were utilized in the reviewed studies are presented in the last column of Table I.

C. COMPARISON OF DIFFERENT ADVERSARIES ON ATTRIBUTES

This section compares different adversarial example generation techniques in terms of attack attributes of capacity, specificity, transferability, and kind of employed perturbation.

1) THE CAPACITY

Table II. summarizes two primary attribute information, i.e., the capacity and the specificity of the attack methods. Regarding the capacity attribute, we found that most of the attack generation techniques are white-box attacks. In the scenario of black-box attacks, focusing on CNN model orientation, Dong *et al.* [72] considered a black-box decision-based attack setting and demonstrated that their approach could converge fast and fool the target model with fine distortions. In [85] an operative black-box adversarial attack was generated against commercial APIs and further step was taken exploring the transferability of feature-level adversarial examples against deep CNN-based FR models (Section IV-B.1). Goodman *et al.* [90] proposed the Advbox toolbox, which showed its ability to support black-box attacks against FR systems. Regarding physical attacks orientation, authors in [78] proposed the *VLA* against black-box FR systems. Nguyen *et al.* [55] focused on real-time light projection-based attacks considering both white- and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

TABLE II
COMPARISON OF DIFFERENT ADVERSARIAL ATTACKS ON CAPACITY AND SPECIFICITY ATTRIBUTES

Representative study	Adversarial capacity	Adversarial Specificity
[71]	None	None
[72]	Black-box	Both
[85]	Black-box	Targeted
[89]	White-box	Non-targeted
[81]	White-box	Targeted
[82]	White-box	Targeted
[90]	Both	Both
[76], [91]	White-box	Both
[92]	White-box	Both
[78]	Black-box	Both
[55]	Both	Both
[80]	White-box	Non-targeted
[93]	White-box	Both
[83]	White-box	Non-targeted
[94]	White-box	Targeted
[56]	Black-box	Both

black-box attack settings. In geometry-oriented attacks, Deb *et al.* [56] demonstrated that faces generated by the *AdvFaces* adversarial face synthesis method could evade several black-box contemporary face-matching techniques while achieving unprecedented attack SRs.

2) THE SPECIFICITY

Considering the specificity of adversarial example generation techniques, Table II. represents that most attack methods are both targeted and non-targeted. Hence, the generalization is practically considered regarding this attribute. In the scenario of non-targeted attacks, which are easier to implement, Garofalo *et al.* [89] concentrated on the poisoning attack design, Komkov and Petiushko [80] focused on the evasion purpose of paper sticker projection on the hats, and Dabouei *et al.* [83] prioritized the speed of their landmark-based adversarial example generation algorithm.

3) THE TRANSFERABILITY

The transferability of attack methods was explored by some studies [56], [80], [85], [91]. Zhong and Deng [85] explored the vulnerability of CNN-based FR models to transferable attacks. They observed that their proposed *DFANet* technique could enhance the transferability of existing attack methods. Sharif *et al.* [91] found that attacks against the OpenFace architecture could successfully fool the VGG architecture in only a limited number of attempts (10–12%), whereas dodging against VGG can lead to successful dodging against OpenFace in at least 63% of attempts. They also argued that the generated universal attacks could transfer between architectures with similar success. Komkov and Petiushko [80] demonstrated that a paper sticker's projection on the hat with their proposed reproducible *AdvHat* method could easily confuse Face ID model LResNet100E-IR. They expressed that the proposed method is transferable to other Face ID models, taken from InsightFace Model Zoo[7], which have different architectures, loss functions, and datasets for training in comparison to the LResNet100E-IR. Deb *et al.* [56] verified that faces generated with their *AdvFaces* adversarial face synthesis method are model-agnostic and transferable and can evade several black-box new face matching techniques.

4) THE PERTURBATION

Though universal perturbations make it easier to create adversaries in real-world applications, all except one reviewed attack methods in this paper have demonstrated to generate image-specific perturbations. In [89], authors generated universal dodging with a small number of eyeglasses that many subjects can use to evade recognition. This is despite the fact that universal perturbation generation against FR models seems to be a potential research path and is worth investing some time to avoid noise reformation any time input samples are altered (Section VI-D).

V. DEFENSE AGAINST ADVERSARIAL EXAMPLES

As novel approaches for crafting adversarial examples are proposed, research is also directed to confront adversaries aiming to moderate their consequence on a target deep network's performance. Accordingly, several defense strategies have been defined to increase the security of at-risk FR models.

A. DEFENSE OBJECTIVES

The objectives of defense strategies could be generally categorized into the following:

Model architecture preservation is a primary consideration when constructing any defense techniques against adversarial examples. With this objective, the minimal alteration should be exerted on model architectures.

Accuracy maintenance is a primary factor considered to keep the classification outputs almost unaffected.

Model speed conservation is another factor that should not be affected during testing with the deployment of defense techniques on large datasets.

B. DEFENSE STRATEGIES

Generally, the defense strategies against the adversarial attacks can be divided into three categories: (1) altering the training during learning, e.g., by injecting adversarial examples into training data or incorporating altered input throughout testing, (2) changing networks, e.g., by changing the number of layers, subnetworks, loss, and activation functions, and (3) supplementing the primary model by external networks to associate in classifying unseen samples. The methodologies in the first category are not concerned with the learning models. However, the other two categories directly deal with the NNs themselves. The difference between 'changing' a network and 'supplementing' a network by external networks is that the former changes the original

[7] https://github.com/deepinsight/insightface/wiki/Model-Zoo

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

deep network architecture/parameters during training. Simultaneously, the latter keeps the original model intact and attaches external model(s) to it in testing. The taxonomy of the described categories is also displayed in Fig. 8. The remainder of this section is organized consistent with this taxonomy.

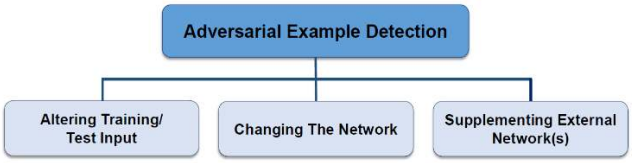

FIGURE 8. A general categorization of adversarial detection methods aimed at defending FR systems against adversarial attacks.

1) ALTERING TRAINING/TEST INPUT

Agarwal *et al.* [13] presented an efficient adversarial detection method to identify an image-agnostic universal perturbation. This method operates on (1) the pixel values and (2) the projections obtained from principal component analysis (PCA) features, as test inputs which are coupled with SVM classifier to detect perturbations. The proposed solution is considered in the first category due to flattening, hence alters the training database's images to form a row vector used either as the pixel values or dimensionally reduced vectors. The authors evaluated the effectiveness of this approach by two perturbation algorithms, universal perturbation, and a variant of it, called fast feature fool [103]. Doing experiments with three different databases, MEDS, PaSC, and Multi-PIE [104], and four different DNN architectures, VGG-16, GoogLeNet, ResNet-152 [45], and CaffeNet [105], they showed that more straightforward approaches, such as the one proposed, can yield higher detection rates for image-agnostic adversarial perturbation. Another research [106] proposed a defense strategy based on an ensemble of classification from domain transformed input data. According to this approach, input images are transformed into a grayscale format, cropped, and rotated to pass the classifier, the predictions of which assembled to create the ensemble decision. The goal of this research was to discover a method that does not necessitate any retraining. On the VGGface2 dataset, experiments showed that domain transformation is useful to suppress the impact of adversarial attacks on face verification tasks.

2) CHANGING THE NETWORK

Goswami *et al.* [14] proposed two defense algorithms: (1) an adversarial perturbation detection algorithm, which utilizes the CNN intermediate filter responses, and (2) a mitigation algorithm, which incorporates a specific dropout technique. In the former, authors compared the patterns of the in-between representations for original images with corresponding distorted images at each layer. They applied the differences of the two patterns to train a classifier that can categorize an unseen input as an original/distorted image. In the latter, they selectively dropped out the most affected filter responses of a CNN model, i.e., filter responses for in-between layers that reflect the most sensitivity towards noisy data to lessen the impact of adversarial noise. Subsequently, they made a comparison with unaffected filter maps. Using the VGG-Face and LightCNN networks, authors assessed the detection and mitigation algorithms according to a cross-database protocol; they performed training only with the Multi-PIE database and accomplished testing MEDS, PaSC, and MBGC [107] databases. Across all distortions on the three databases, it was shown that the proposed detection algorithm maintains high true-positive rates even at low false-positive rates, which are desirable for the system. Also, it was observed that by discarding a certain fraction of the most affected in-between representations with the proposed mitigation algorithm, better recognition outputs could be achieved.

In another study, a blockchain security mechanism is presented to protect against FR models' attacks [108] presented. Traditional blocks of any deep learning models, such as CNNs, are converted into blocks similar to the blockchain blocks to offer fault-tolerant access in a distributed setting. In this way, tampering in one specific component alerts the entire system and easily detects 'any' probable alteration. Experiments revealed the proposed network's resilience to both the deep learning model and the biometric template, using Multi-PIE and MEDS databases.

Su *et al.* [109] proposed a deep *Residual Generative Network* (*ResGN*) to clean adversarial perturbations for face verification. They suggested an innovative training framework composed of *ResGN*, VGG-Face, and FaceNet; they presented a joint of three losses: a pixel loss, a texture loss, and a verification loss, to optimize *ResGN* parameters. The VGG-Face and FaceNet networks contribute to the learning procedure by providing texture and verification losses, respectively, hence, improve the verification performance of cleaned images fundamentally. The empirical results validated the effectiveness of the proposed method on the LFW benchmark dataset. Zhong and Deng [75] offered to recover the local smoothness of the representation space by integrating a margin-based triplet embedding regularization (MTER) term into the classification objective so that the acquired model learns to resist adversarial examples. The regularization term consists of a two-phase optimization that detects probable perturbations and punishes those using a large margin in an iterative approach. Experimental outcomes on CASIA-WebFace, VGGFace2, and MS-Celeb-1M demonstrated that the proposed method elevates network robustness against both feature-level and label-level adversarial attacks in deep FR models.

According to the concept of feature distance spaces explored in [110], Massoli *et al.* [111] proposed a detection approach based on the trajectory of internal representations, i.e., hidden layers' neuron activation, also known as deep

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

features. They argued that the representations of adversarial inputs follow a different evolution for genuine inputs. Specifically, they collected deep features during the forward step of the target model, applied average pooling over deep features to achieve a single features vector at each selected layer, and computed the distance between each vector and the class centroid of each class at each layer, to acquire an embedding that represents the trajectory of the input image in the features space. Such a trajectory was finally fed to a binary classifier or adversarial detector. As the adversarial detector, two different architectures of a multilayer perceptron (MLP) and a long-short term memory (LSTM) network were considered in this work. The authors conducted the experiments on the VGGFace2 dataset and the state-of-the-art Se-ResNet-50 [52]. To assess the efficiency of the proposed approach, they showed the receiving operating characteristics (ROC) curves from the adversarial detection considering targeted and non-targeted attacks for each architecture. They reported the area under the curve (AUC) values relative to each attack. Accordingly, the AUC values were very close for the targeted attacks, whereas, in the case of non-targeted attacks, the LSTM performance was shown to be considerably better than the MLP.

Recently, Kim *et al.* [112] proposed a low-power, highly secure always-on FR processor for verification applications on mobile devices. This processor operates based on three key features of (1) a branch net-based early stopping FR (BESF) method to prevent adversarial attacks and consume low power, (2) a unified processing element (PE) for point- and depth-wise convolutions with layer fusion to reduce external memory access and (3) a noise injection layer (NIL) incorporated between bottleneck layers to make the network more robust against adversarial attacks with lower external memory access. They demonstrated that under the FGSM and PGD, BESF could result in high recognition accuracies while reducing the average power consumption significantly. They also showed that the PE reduces the external memory access, and the NIL could further lessen the FGSM and PGD attack SRs. Overall, this processor resulted in 95.5% FR accuracy in the Labeled Faces in the LFW dataset.

3) SUPPLEMENTING EXTERNAL NETWORK

Xu *et al.* [113] proposed a feature squeezing strategy that moderates the search space available to an adversary by coalescing samples correspond to different feature vectors in the original space into a single sample. Adding two external models to the classifier network, they explored two feature squeezing approaches by (1) decreasing the color bit depth of each pixel and (2) spatial smoothing. Goswami *et al.* [14] expressed that this approach is simple and operative for high-resolution images with detailed data; however, it may not be operational for low resolution cropped faces frequently used in FR settings. In [114], an open-source Python-based toolbox, termed as SmartBox, is proposed to benchmark the function of adversarial attack detection and mitigation algorithms against FR models. The detection approaches included in this toolbox are: 'Detection via Convolution Filter Statistics,' 'PCA-based detection,' 'Artifacts Learning' and 'Adaptive' Noise Reduction,' which are respectively considered in 'Changing the Network,' 'Altering Training/Test Input,' and 'Supplementing External Networks' defense categories. We put this study under the 'Supplementing External Networks' category since it covers the last two and hence, the majority of SmartBox detection methods.

While most of the current defense methods either assume prior knowledge of specific attacks or may not operate well on complex models due to their underlying assumptions, a new window was opened to adversarial detection techniques by leveraging the interpretability of DNNs [15]. Tao *et al.* [15] proposed a detection technique called *Attacks meet Interpretability* (*AmI*) in the context of FR practice. This technique features an innovative bi-directional correspondence inference amongst face attributes and internal neurons, using attribute-level mutation and neuron strengthening/weakening. More precisely, critical neurons for individual attributes are identified, and the activation values are enhanced to amplify the reasoning part of the computation. In contrast, other neurons' activation values are weakened to suppress the uninterpretable part. Employing three different datasets, VGG-Face, LFW, and CelebA, *AmI* applied to VGG-Face, with seven different kinds of attack. Extensive experiments represented that the proposed technique could successfully detect adversarial samples with a true-positive rate of 94% on average, which is significantly higher than what was achieved with the state-of-the-art reference technique, called feature squeezing [113]. Similarly, the FPR of the *AmI* technique, is lower than the reference work, demonstrating its high effectiveness in this endeavor. A general overview of different adversarial example detection approaches, along with their category, is provided in Table III.

VI. CHALLENGES AND DISCUSSION

Although several adversarial example generation methods and defense strategies have been proposed and developed in FR's realm, various problems and challenges need to be addressed. This section summarizes the potential challenges that threaten this field. We categorize the challenges into four groups based on the literature reviewed above.

A. PARTICULARIZATION/SPECIFICATION OF ADVERSARIAL EXAMPLES

As described in this study, several image-, face-, and feature-level adversarial example generation methods have been proposed to fool FR systems; however, these methods are challenging to construct a generalized adversarial example and can only achieve good performance in a certain evaluation metrics. These evaluation metrics are mainly divided into three categories: The SR to generate adversarial examples, the robustness of the FR models, and specific attributes of the attacks, such as the perturbation amount and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092646, IEEE Access

TABLE III
ADVERSARIAL EXAMPLE DETECTION APPROACHES

Representative study	Defense strategies	Description
[13]	Altering training/test input	Image pixels + PCA + SVM
[106]	Altering training/test input	An ensemble of classification results from domain transformed (grayscale, cropped and rotated) input data
[14]	Changing the network	Filter responses of CNN; dropout of filter responses
[108]	Changing the network	Conversion of traditional blocks of deep learning models into blocks similar to the blocks in the blockchain
[109]	Changing the network	Design of *ResGN* model + employment of a pixel loss, a texture loss, and a verification loss for parameter optimization
[75]	Changing the network	Integration of MTER term into the classification objective for detection and punishment of perturbations
[111]	Changing the network	Exploration of the adversary's evolution by tracking the trajectory of deep features representations
[112]	Changing the network	Design of a low-power and highly secure always-on FR processor
[113]	Supplementing external network(s)	Feature squeezing strategies of (1) pixel's color bit depth decreasing and (2) spatial smoothing via the addition of two external models to the classifier
[114]	Supplementing external network(s)	SmartBox toolbox
[15]	Supplementing external network(s)	Bi-directional correspondence inference amongst face attributes and internal neurons via *AmI* technique
[115]	Supplementing external network(s)	Defending black-box FR classifiers via iterative adversarial image purifiers

degree of the transferability. To explain briefly, the SR of an attack, known as the most direct and effective evaluation criterion, is inversely proportional to the magnitude of perturbations. The robustness of FR models is related to the classification accuracy. The better the design of the FR model, the less it is vulnerable to adversarial examples. Regarding the attacks' attributes, too small perturbations on the original examples are difficult to construct adversarial examples, whereas too large perturbations are easily distinguished by human eyes. Therefore, a balance between constructing adversarial examples and the human visual system should be achieved. On the other hand, within a certain perturbation range, the transfer rate of adversarial examples is proportional to the magnitude of adversarial perturbations, i.e., the greater perturbations to the original example, the higher the transfer rate of the constructed adversarial examples. Considering these facts, the amount of perturbation to be considered on the original images, and the design of model architecture becomes critical.

Similarly, the variations in imaging conditions investigated in different works are narrower than can be encountered in practice. i.e., they are happened to be in controlled lighting, distance, etc. These conditions could be applied to some practical cases (e.g., an FR system deployed within a building). However, other practical scenarios are more challenging, needing the attacks to be tolerant of a more extensive range of imaging conditions.

These matters inhibit the defenders from designing generalized detection techniques and encourage them to propose efficient defenses against confined attacks. To overcome such challenges, a comprehensive experimental setup should be considered, possibly via scheming a standard platform as a benchmark setup setting, so that all evaluation metrics are measured simultaneously to report the efficiency of generated adversarial examples. Also, the research space should be focused more on (1) the amount of perturbation to be considered on the original images, (2) the design of FR models' architectures to be targeted, and (3) the level of transferability of generated adversarial examples. As demonstrated in Table II, the vulnerability of existing FR models to adversarial attacks in a black-box manner has been studied less, revealing the lack of transferability exploration.

B. INSTABILITY OF FR MODELS

Though the introduction of deep FR systems has brought benefits, it has increased the attack surface of such systems. Implementing image distortion-based adversarial attacks, for example, a substantial loss in the performance of deep learning-based systems observed, compared with the application of shallow learning-based commercial-off-the-shelf matchers for the same evaluation data. Accordingly, the integration of only those architectures that are robust against evasion is strongly advocated. The need to develop robust models to increase adversarial examples' generalizability has been expressed in the previous paragraph, along with other influencing factors. However, this obligation is restated separately to emphasize its importance when taking steps toward generating more black-box attacks. In these circumstances, security concerns for developing more robust FR models will be raised.

C. DEVIATION FROM THE HUMAN VISION SYSTEM

Adversarial attacks on vision systems exploit the fact that systems are sensitive to small changes in images to which humans are not. It will be a good idea to develop algorithms that reason images more similar to humans. In particular, those approaches that classify images based on their attributes rather than on their pixels' intensities may become more practical. Such approaches may train classifiers to recognize the presence or absence of describable aspects of visual appearances, like gender, race, age, and hair color, and extract and compare high-level visual features, or traits, of a face image that are insensitive to pose, illumination, expression, and other imaging conditions.

Profound regard to human vision physiology may open another window to research space as well. For example, the *VLA* manifested a successful implementation of physical adversarial attacks, in the design of which an attempt was made to emulate the human visual system.

D. IMAGE-AGNOSTIC PERTURBATION GENERATION

The existing adversarial example generation methods are remarkably demonstrated to be image-agnostic, and the lack of universal perturbation generation against FR models is strongly noticed. An FR model's capability to attack different target faces simultaneously would be the by-product of generating universal perturbations, which is an essential concern in numerous studies that have been conducted in this regard.

VII. CONCLUSION

This article presented a comprehensive survey in the course of adversarial attacks against intelligent deep FR systems. Despite the outstanding performance of advanced FR models, they have been vulnerable to imperceptible adversarial input images that lead them to modify their outputs entirely. This fact has opened a new window to numerous recent contributions to devise adversarial attacks and countermeasures in the FR systems. This article reviewed these contributions, mainly concentrating on the most effective and inspiring works in the literature. A taxonomy of existing attack and defense methods is proposed based on different criteria. We also discussed current challenges and potential solutions in adversarial examples targeting FR models. Hope this work can shed some light on the key concepts to encourage progress in the future.

REFERENCES

[1] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," 1991, pp. 586–591, doi: 10.5120/20740-3119.

[2] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[4] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*, 2016, pp. 499–515.

[5] R. Singh, A. Agarwal, M. Singh, S. Nagpal, and M. Vatsa, "On the robustness of face recognition algorithms against attacks and bias," *arXiv Prepr. arXiv2002.02942*, 2020.

[6] S. Marcel, M. S. Nixon, and S. Z. Li, *Handbook of biometric anti-spoofing*, vol. 1. Springer, 2014.

[7] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Comput. Surv.*, vol. 50, no. 1, pp. 1–37, 2017.

[8] C. Rathgeb, P. Drozdowski, and C. Busch, "Makeup presentation attacks: Review and detection performance benchmark," *IEEE Access*, 2020.

[9] C. Szegedy *et al.*, "Intriguing properties of neural networks," *arXiv Prepr. arXiv1312.6199*, 2013.

[10] M. Ferrara, A. Franco, and D. Maltoni, "The magic passport," in *IEEE International Joint Conference on Biometrics*, 2014, pp. 1–7.

[11] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. neural networks Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, 2019.

[12] H. Xu *et al.*, "Adversarial attacks and defenses in images, graphs and text: A review," *Int. J. Autom. Comput.*, vol. 17, no. 2, pp. 151–178, 2020.

[13] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha, "Are image-agnostic universal adversarial perturbations for face recognition difficult to detect?," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018, pp. 1–7.

[14] G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa, "Detecting and mitigating adversarial perturbations for robust face recognition," *Int. J. Comput. Vis.*, vol. 127, no. 6–7, pp. 719–742, 2019.

[15] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 7717–7728.

[16] L. Guarnera, O. Giudice, and S. Battiato, "Fighting Deepfake by Exposing the Convolutional Traces on Images," *IEEE Access*, vol. 8, pp. 165085–165098, 2020.

[17] Y. Zhou, X. Hu, L. Wang, S. Duan, and Y. Chen, "Markov chain based efficient defense against adversarial examples in computer vision," *IEEE Access*, vol. 7, pp. 5695–5706, 2018.

[18] Y. Bakhti, S. A. Fezza, W. Hamidouche, and O. Déforges, "Ddsa: a defense against adversarial attacks using deep denoising sparse autoencoder," *IEEE Access*, vol. 7, pp. 160397–160407, 2019.

[19] X. Liu *et al.*, "Privacy and Security Issues in Deep Learning: A Survey," *IEEE Access*, 2020.

[20] A. Makrushin, T. Neubert, and J. Dittmann, "Automatic Generation and Detection of Visually Faultless Facial Morphs," in *VISIGRAPP (6: VISAPP)*, 2017, pp. 39–50.

[21] D. J. Robertson, R. S. S. Kramer, and A. M. Burton, "Fraudulent ID using face morphs: Experiments on human and automatic recognition," *PLoS One*, vol. 12, no. 3, p. e0173319, 2017.

[22] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch, "Transferable deep-cnn features for detecting digital and print-scanned morphed face images," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1822–1830.

[23] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, "Detection of face morphing attacks by deep learning," in *International Workshop on Digital Watermarking*, 2017, pp. 107–120.

[24] L. Debiasi, U. Scherhag, C. Rathgeb, A. Uhl, and C. Busch, "PRNU-based detection of morphed face images," in *2018 International Workshop on Biometrics and Forensics (IWBF)*, 2018, pp. 1–7.

[25] L.-B. Zhang, F. Peng, and M. Long, "Face morphing detection using Fourier spectrum of sensor pattern noise," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[26] M. Ferrara, A. Franco, and D. Maltoni, "Face demorphing," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 4, pp. 1008–1017, 2017.

[27] F. Peng, L.-B. Zhang, and M. Long, "FD-GAN: Face de-morphing generative adversarial network for restoring accomplice's facial image," *IEEE Access*, vol. 7, pp. 75122–75131, 2019.

[28] D. Ortega-Delcampo, C. Conde, D. Palacios-Alonso, and E. Cabello, "Border control morphing attack detection with a convolutional neural network de-morphing approach," *IEEE Access*, vol. 8, pp. 92301–92313, 2020.

[29] N. Carlini, "A Complete List of All (arXiv) Adversarial Example Papers," 2019. https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html.

[30] M. Wang and W. Deng, "Deep Face Recognition: A Survey," *arXiv*, p. arXiv-1804, 2018.

[31] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1864–1870, 2012.

[32] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, 2005.

[33] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond eigenfaces: Probabilistic matching for face recognition," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 30–35.

[34] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?," in *2011 International conference on computer vision*, 2011, pp. 471–478.

[35] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, 2002.

[36] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, 2006.

[37] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *2010 IEEE Computer society conference on computer vision and pattern recognition*, 2010, pp. 2707–2714.

[38] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?," *IEEE Trans. image Process.*, vol. 24, no. 12, pp. 5017–5032, 2015.

[39] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 289–302, 2013.

[40] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," 2008.

[41] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[42] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv Prepr. arXiv1409.1556*, 2014.

[44] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

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

[46] H. Wang *et al.*, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[47] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[48] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 11, pp. 2884–2896, 2018.

[49] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[50] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv Prepr. arXiv1411.7923*, 2014.

[51] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*, 2016, pp. 87–102.

[52] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 67–74.

[53] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4873–4882.

[54] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv Prepr. arXiv1607.02533*, 2016.

[55] D.-L. Nguyen, S. S. Arora, Y. Wu, and H. Yang, "Adversarial Light Projection Attacks on Face Recognition Systems: A Feasibility Study," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 814–815.

[56] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face

synthesis," *arXiv Prepr. arXiv1908.05008*, 2019.

[57] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[58] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," *C. Sch. Comput. Sci.*, vol. 6, p. 2, 2016.

[59] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv Prepr. arXiv1412.6572*, 2014.

[60] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, 2018.

[61] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv Prepr. arXiv1611.01236*, 2016.

[62] A. Rozsa, E. M. Rudd, and T. E. Boult, "Adversarial diversity and hard positive generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 25–32.

[63] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*, 2016, pp. 372–387.

[64] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, 2019.

[65] S. Das and P. N. Suganthan, "Differential evolution: A survey of the state-of-the-art," *IEEE Trans. Evol. Comput.*, vol. 15, no. 1, pp. 4–31, 2010.

[66] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[67] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[68] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, 2017, pp. 39–57.

[69] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv Prepr. arXiv1503.02531*, 2015.

[70] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*, 2016, pp. 582–597.

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

[72] Y. Dong *et al.*, "Efficient decision-based black-box adversarial attacks on face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7714–7722.

[73] C. Igel, T. Suttorp, and N. Hansen, "A computational efficient covariance matrix update and a (1+ 1)-CMA for evolution strategies," in *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, 2006, pp. 453–460.

[74] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evol. Comput.*, vol. 9, no. 2, pp. 159–195, 2001.

[75] Y. Zhong and W. Deng, "Adversarial Learning with Margin-based Triplet Embedding Regularization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6549–6558.

[76] 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 *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016, pp. 1528–1540.

[77] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

[78] 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," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–19, 2019.

[79] L. Zhang *et al.*, "Kaleido: You can watch it but cannot record it," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 372–385.

[80] S. Komkov and A. Petiushko, "AdvHat: Real-world adversarial attack on ArcFace Face ID system," *arXiv Prepr. arXiv1908.08705*, 2019.

[81] E. Chatzikyriakidis, C. Papaioannidis, and I. Pitas, "Adversarial Face De-Identification," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 684–688.

[82] H. Kwon, O. Kwon, H. Yoon, and K.-W. Park, "Face Friend-Safe Adversarial Example on Face Recognition System," in *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*, 2019, pp. 547–551.

[83] A. Dabouei, S. Soleymani, J. Dawson, and N. Nasrabadi, "Fast geometrically-perturbed adversarial faces," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 1979–1988.

[84] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," *arXiv Prepr. arXiv1801.02612*, 2018.

[85] Y. Zhong and W. Deng, "Towards Transferable Adversarial Attack against Deep Face Recognition," *arXiv Prepr. arXiv2004.05790*, 2020.

[86] Y. Dong *et al.*, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[87] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv Prepr. arXiv1611.02770*, 2016.

[88] C. Xie *et al.*, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2730–

2739.

[89] G. Garofalo, V. Rimmer, D. Preuveneers, and W. Joosen, "Fishy faces: crafting adversarial images to poison face authentication," 2018.

[90] D. Goodman, H. Xin, W. Yang, W. Yuesheng, X. Junfeng, and Z. Huan, "Advbox: a toolbox to generate adversarial examples that fool neural networks," *arXiv Prepr. arXiv2001.05574*, 2020.

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

[92] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang, "Invisible mask: Practical attacks on face recognition with infrared," *arXiv Prepr. arXiv1803.04683*, 2018.

[93] M. Pautov, G. Melnikov, E. Kaziakhmedov, K. Kireev, and A. Petiushko, "On adversarial patches: real-world attack on ArcFace-100 face recognition system," in *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 2019, pp. 391–396.

[94] Q. Song, Y. Wu, and L. Yang, "Attacks on state-of-the-art face recognition using attentional adversarial attack generative network," *arXiv Prepr. arXiv1811.12026*, 2018.

[95] A. Majumdar, R. Singh, and M. Vatsa, "Face verification via class sparsity based supervised encoding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1273–1280, 2016.

[96] J. R. Beveridge *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

[97] A. P. Founds, N. Orlans, W. Genevieve, and C. I. Watson, "Nist special databse 32-multiple encounter dataset ii (meds-ii)," 2011.

[98] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *arXiv Prepr. arXiv1712.04248*, 2017.

[99] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," *arXiv Prepr. arXiv1807.04457*, 2018.

[100] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," *arXiv Prepr. arXiv1804.08598*, 2018.

[101] F. Wang *et al.*, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 765–780.

[102] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *2014 IEEE international conference on image processing (ICIP)*, 2014, pp. 343–347.

[103] K. R. Mopuri, U. Garg, and R. V. Babu, "Fast feature fool: A data independent approach to universal adversarial perturbations," *arXiv Prepr. arXiv1707.05572*, 2017.

[104] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[105] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 675–678.

[106] L. Kurnianggoro and K.-H. Jo, "Ensemble of Predictions from Augmented Input as Adversarial Defense for Face Verification System," in *Asian Conference on Intelligent Information and Database Systems*, 2019, pp. 658–669.

[107] P. J. Phillips *et al.*, "Overview of the multiple biometrics grand challenge," in *International Conference on Biometrics*, 2009, pp. 705–714.

[108] A. Goel, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha, "Securing CNN model and biometric template using blockchain," *IEEE BTAS*, 2019.

[109] Y. Su, G. Sun, W. Fan, X. Lu, and Z. Liu, "Cleaning adversarial perturbations via residual generative network for face verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2597–2601.

[110] F. Carrara, R. Becarelli, R. Caldelli, F. Falchi, and G. Amato, "Adversarial examples detection in features distance spaces," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, p. 0.

[111] F. V. Massoli, F. Carrara, G. Amato, and F. Falchi, "Detection of Face Recognition Adversarial Attacks," *arXiv Prepr. arXiv1912.02918*, 2019.

[112] Y. Kim, D. Han, C. Kim, and H.-J. Yoo, "A 0.22–0.89 mW Low-Power and Highly-Secure Always-On Face Recognition Processor With Adversarial Attack Prevention," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 67, no. 5, pp. 846–850, 2020.

[113] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv Prepr. arXiv1704.01155*, 2017.

[114] A. Goel, A. Singh, A. Agarwal, M. Vatsa, and R. Singh, "Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018, pp. 1–7.

[115] R. Theagarajan and B. Bhanu, "Defending Black Box Facial Recognition Classifiers Against Adversarial Attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 812–813.