FACESEC: A Fine-grained Robustness Evaluation Framework for Face Recognition Systems

Liang Tong[1,2*], Zhengzhang Chen[2†], Jingchao Ni[2], Wei Cheng[2],
Dongjin Song[3], Haifeng Chen[2], Yevgeniy Vorobeychik[1]
[1]Washington University in St. Louis, {liangtong, yvorobeychik}@wustl.edu
[2]NEC Laboratories America, {zchen,jni,weicheng,haifeng}@nec-labs.com
[3]University of Connecticut, dongjin.song@uconn.edu

Abstract

We present FACESEC*, a framework for fine-grained robustness evaluation of face recognition systems.* FACESEC *evaluation is performed along four dimensions of adversarial modeling: the nature of perturbation (e.g., pixel-level or face accessories), the attacker's system knowledge (about training data and learning architecture), goals (dodging or impersonation), and capability (tailored to individual inputs or across sets of these). We use* FACESEC *to study five face recognition systems in both closed-set and open-set settings, and to evaluate the state-of-the-art approach for defending against physically realizable attacks on these. We find that accurate knowledge of neural architecture is significantly more important than knowledge of the training data in black-box attacks. Moreover, we observe that open-set face recognition systems are more vulnerable than closed-set systems under different types of attacks. The efficacy of attacks for other threat model variations, however, appears highly dependent on both the nature of perturbation and the neural network architecture. For example, attacks that involve adversarial face masks are usually more potent, even against adversarially trained models, and the ArcFace architecture tends to be more robust than the others.*

1. Introduction

Face recognition has received much attention [12, 23, 25, 34, 15, 33] in recent years. Empowered by deep convolutional neural networks (CNNs), it has become widely used in various areas, including security-sensitive applications, such as airport check-in, online financial transactions, and mobile device login. The success of such *deep face recognition* is particularly striking, with >99% prediction accuracy on benchmark datasets [23, 16, 15, 6].

Despite its widespread success in computer vision applications, recent studies have found that deep face recognition models are vulnerable to *adversarial examples* in both *digital space* [18, 8, 36] and *physical space* [26]. The former directly modifies an input face image by adding imperceptible perturbations to mislead face recognition (henceforth, *digital attacks*). The latter is characterized by adding adversarial perturbations that can be realized on *physical objects* (*e.g.*, wearing an adversarial eyeglass frame [26]), which are subsequently captured by a camera and then fed into a face recognition model to fool prediction (henceforth, *physically realizable attacks*). As such, the aforementioned domains, especially critical domains such as security and finance, are subjected to risks of opening the backdoor for the attackers. For example, in face recognition supported financial/banking services, an illegal user may bypass biometric verification and steal money from victims' accounts. Therefore, there exists a vital need for methods that can comprehensively and systematically evaluate the robustness of face recognition systems in adversarial settings, which in turn can shed light on the design of robust models for downstream tasks.

The main challenges of comprehensive evaluation of the robustness of face recognition lie in dealing with the diversity of face recognition systems and adversarial environments. First, different face recognition systems consist of various key components (*e.g.*, training data and neural architecture); such diversity results in different performance and robustness. To enable comprehensive and systematic evaluations, it is crucial to assess the robustness of every individual or a combination of face recognition components in adversarial settings. Second, adversarial example attacks can vary by the nature of perturbations (*e.g.*, pixel-level or physical space), an attacker's goal, knowledge, and capability. For a given face recognition system, its robustness against a specific type of attack may not generalize to other kinds [35].

In spite of recent advances in adversarial attacks [26, 8, 36] that demonstrate the vulnerability of face recognition

*Work done during an internship at NEC Laboratories America.
†Corresponding author.

systems, most existing methods fail to address the aforementioned challenges due to the following reasons. First, current efforts appeal to either *white-box attacks* or *black-box attacks* to obtain a lower bound or upper bound of robustness. These bounds indicate the vulnerability of face recognition systems in adversarial settings but lack the understanding of how each component of face recognition contributes to such vulnerability. Second, while most existing approaches focus on a specific type of attack (*e.g.*, digital attacks that incur imperceptible noise [8, 36]), they fail to explore the different levels of robustness in response to various attacks (*e.g.*, physically realizable attacks).

To bridge this gap, we propose FACESEC, a fine-grained robustness evaluation framework for face recognition systems. FACESEC incorporates four dimensions in evaluation: the nature of adversarial perturbations (pixel-level or face accessories), the attacker's accurate knowledge about the target face recognition system (training data and neural architecture), goals (dodging or impersonation), and capability (individual or universal attacks). Specifically, we implement both digital and physically realizable attacks in FACESEC. We leverage the PGD attack [18], the state-of-the-art digital attack paradigm, and the eyeglass frame attack [26] as the representative of physically realizable attacks. Additionally, we propose two novel physically realizable attacks: one involves pixel-level adversarial stickers on human faces, and the other adds color grids on face masks. Moreover, to facilitate universal attacks that produce *image-agnostic* perturbations, we propose a systematic approach that works on top of the attack paradigms described above.

In summary, this paper makes the following contributions:

(1) We propose FACESEC, the first robustness evaluation framework that enables researchers to (i) identify the vulnerability of each face recognition component to adversarial examples, and (ii) assess different levels of robustness under various adversarial circumstances.

(2) We propose two novel physically realizable attacks: the pixel-level sticker attack and the grid-level face mask attack. These allow us to explore adversarial robustness against different types of physically realizable perturbations. Particularly, the latter responds to the pressing needs for security analysis of face recognition systems, as face masks have become common face accessories during the COVID-19 pandemic.

(3) We propose a general approach to produce universal adversarial examples for a batch of face images. Compared to previous works, our paradigm has a significant speedup and is more efficient in evaluation.

(4) We perform a comprehensive evaluation on five publicly available face recognition systems in various settings to demonstrate the efficacy of FACESEC.

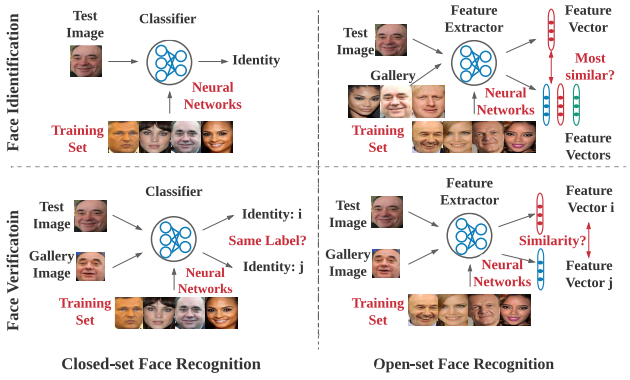

Figure 1. Closed-set and open-set face recognition systems.

2. Background and Related Work

2.1. Face Recognition Systems

Generally, deep face recognition systems aim to solve the following two tasks: 1) *Face identification*, which returns the predicted identity of a test face image; 2) *Face verification*, which indicates whether a test face image (also called probe face image) and the face image stored in the gallery belong to the same identity. Based on whether all testing identities are predefined in the training set, face recognition systems can be further categorized into *closed-set systems* and *open-set systems* [15], as illustrated in Fig. 1.

In closed-set face recognition tasks, all the testing samples' identities are enrolled in the training set. Specifically, a face identification task is equivalent to a *multi-class classification* problem by using the standard softmax loss function in the training phase [31, 28, 27]. And a face verification task is a natural extension of face identification by first performing the classification twice (one for the test image and the other for the gallery) and then comparing the predicted identities to see if they are identical.

In contrast, there are usually no overlaps between identities in the training and testing set for open-set tasks. In this setting, a face verification task is essentially a *metric learning* problem, which aims to maximize *intra-class distance* and minimize *inter-class distance* under a chosen metric space by two steps [25, 23, 34, 16, 15, 6]. First, we train a feature extractor that maps a face image into a discriminative feature space by using a carefully designed loss function; Then, we measure the distance between feature vectors of the test and gallery face images to see if it is above a verification threshold. As an extension of face verification, the face identification task requires additional steps to compare the distances between the feature vectors of the test image and each gallery image, and then choose the gallery's identity corresponding to the shortest distance.

This paper focuses on face identification for closed-set systems, as face verification is just an extension of identification in this setting. Likewise, we focus on face verification

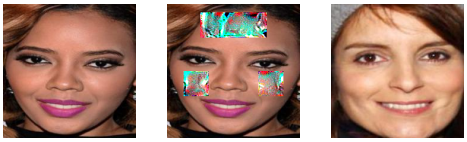

Figure 2. Sticker attack: an example of physically realizable attacks on face recognition systems. Left: original input image. Middle: adversarial sticker on the face. Right: predicted identity. In practice, the adversarial stickers can be printed and put on human faces.

for open-set systems.

2.2. Digital and Physical Adversarial Attacks

Recent studies have shown that deep neural networks are vulnerable to adversarial attacks. These attacks produce *imperceptible* perturbations on images in the digital space to mislead classification [30, 9, 4] (henceforth, *digital attacks*). While a number of attacks on face recognition fall into this category (*e.g.*, by adding small ℓ_p bounded noise over the entire input [8] or perceptible but semantically meaningful perturbation on a restricted area of the input [24]), of particular interest in face recognition, are attacks in the physical world (henceforth, *physical attacks*).

Generally, physical attacks have three characteristics [35]. First, the attackers directly modify the actual entity rather than digital features. Second, the attacks can mislead state-of-the-art face recognition systems. Third, the attacks have low suspiciousness (*i.e.*, by adding objects similar to common "noise" on a small part of human faces). For example, an attacker can fool a face recognition system by wearing an adversarial eyeglass frame [26], a standard face accessory in the real world.

In this paper, we focus on both digital attacks and *the digital representation of physical attacks* (henceforth, *physically realizable attacks*). Specifically, physically realizable attacks are digital attacks that can produce adversarial perturbations with low suspiciousness, and these perturbations can be realized in the physical world by using techniques such as 3-D printing (*e.g.*, Fig. 2 illustrates one example of such attacks on face recognition systems). Compared to physical attacks, physically realizable attacks can evaluate robustness of face recognition systems more efficiently: on the one hand, realizable attacks allow us to iteratively modify digital images directly so the evaluation can significantly speedup compared to modifying real-world objects and then photographing them; on the other hand, robustness to physically realizable attacks provides the lower bound of robustness to physical attacks, as the former has fewer constraints and larger solution space.

Formally, both digital and physically realizable attacks can be performed by solving the following general form of an optimization problem (*e.g.*, for closed-set identification task):

$$\arg\max_{\boldsymbol{\delta}} \ell(S(\boldsymbol{x} + M\boldsymbol{\delta}), y) \qquad s.t.\ \boldsymbol{\delta} \in \Delta, \tag{1}$$

where S is the target face recognition model, ℓ is the adversary's utility function (*e.g.*, the loss function used to train S), $\boldsymbol{x}$ is the original input face image, y is the associated identity, $\boldsymbol{\delta}$ is the adversarial perturbation, and Δ is the feasible space of the perturbation. Here, M denotes the mask matrix that constrains the area of perturbation; it has the same dimension as $\boldsymbol{\delta}$ and contains 1s where perturbation is allowed, and 0s where there is no perturbation.

2.3. Adversarial Defense for Face Recognition

While there have been numerous defense approaches to make face recognition robust to adversarial attacks, many of them focus on digital attacks and have been proved to be broken under adaptive attacks [4, 32]. Here, we describe one representative defense approach, *adversarial training* [18], that is scalable, not defeated by adaptive attacks, and has been leveraged to defend against physically realizable attacks on face recognition systems.

The main idea of adversarial training is to minimize prediction loss of the training data, where an attacker tries to maximize the loss. In practice, this can be done by iteratively using the following two steps: 1) Use an attack method to produce adversarial examples of the training data; 2) Use any optimizer to minimize the loss of predictions on these adversarial examples. Wu *et al.* [35] propose to use DOA—adversarial training with the rectangular occlusion attacks—to defend against physically realizable attacks on closed-set face recognition systems. Specifically, the rectangular occlusion attack included in DOA first heuristically locates a rectangular area among a collection of possible regions in an input face image, then fixes the position and adds adversarial occlusion inside the rectangle. It has been shown that DOA can significantly improve the robustness against the eyeglass frame attack [26] for closed-set VGG-based face recognition system [23] by 80%. However, as we will show in Section 4, DOA would fail to defend against other types of attacks, such as the face mask attack proposed in Section 3.1.

3. Methodology

In this section, we introduce FACESEC for fine-grained robustness evaluation of face recognition systems. Our goal is twofold: 1) identify vulnerability/robustness of each essential component that comprises a face recognition system, and 2) assess robustness in a variety of adversarial settings. Fig. 3 illustrates an overview of FACESEC. Let $S = f(h; D)$ be a face recognition system with a neural architecture h that is trained on a training set D by an algorithm f (*e.g.*, stochastic gradient descent), FACESEC evaluates the robustness of S via a quadruplet:

$$\mathsf{Robustness} = \mathsf{Evaluate}(S, < P, K, G, C >), \tag{2}$$

where $< P, K, G, C >$ represents an attacker who tries to produce adversarial examples to fool S. P is the pertur-

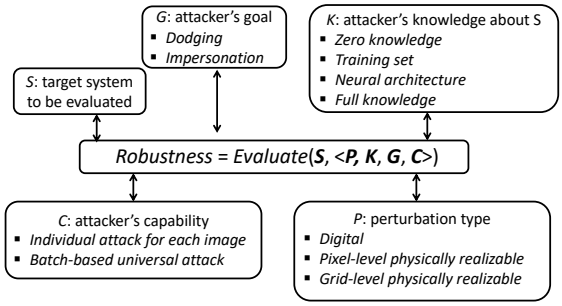

Figure 3. An overview of FACESEC.

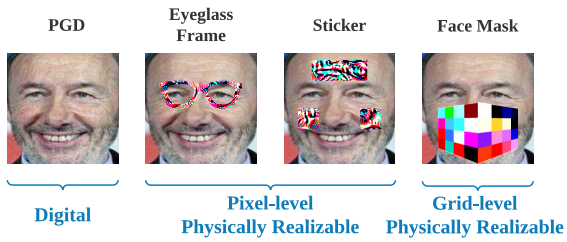

Figure 4. Perturbation types in FACESEC.

bation type, such as perturbations produced by pixel-level digital attacks and physically realizable attacks. K denotes the attacker's knowledge on the target system S, *i.e.*, the information about which sub-components of S are leaked to the attacker. G is the goal of the attacker, such as the circumvention of detection and the misrecognition as a target identity. C represents the attacker's capability. For example, an attacker can either individually perturb each input face image, or produce universal perturbations for images batch-wise. Next, we will describe each element of FACESEC in details.

3.1. Perturbation Type (P)

In FACESEC, we consider three categories of attacks with different perturbation types: *digital attack*, *pixel-level physically realizable attack*, and *grid-level physically realizable attack*, as shown in Fig. 4.

Digital Attack. Digital attack produces small perturbations on the entire input face image. We use the ℓ_∞-norm version of the PGD attack [18] as the representative of this category[1].

Pixel-level Physically Realizable Attack. This category of attack features pixel-level perturbations that can be realized in the physical world (*e.g.*, by printing them on glossy photo papers). In this case, the attacker adds large pixel-level perturbations on a small area of the input image (*e.g.*, face accessories). In FACESEC, we use two attacks of this category: *eyeglass frame attack* [26] and *sticker attack*. The

[1]We also tried other digital attacks (*e.g.*, CW [4] and JSMA [22]), but these were either less effective than PGD or unable to be extended to universal attacks (see Section 3.4).

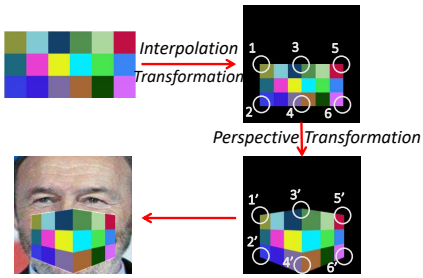

Figure 5. Transformations for the grid-level face mask attack.

former allows large perturbations within an eyeglass frame, and it can successfully mislead VGG-based face recognition systems [23]. We propose the latter to produce pixel-level perturbations that are added on less important face areas than the eyeglass frame, *i.e.*, the two cheeks and forehead of human faces, as illustrated in Fig. 2 and 4. Typically, the stickers are rectangular occlusions, which cover a total of about 20% area of an input face image.

Grid-level Physically Realizable Attack. In practice, pixel-level perturbations are not printable on face accessories made of *coarse* materials, such as face masks using cloths and non-woven fabrics. To address this issue, we propose the grid-level physically realizable face mask attack, which adds a color grid on face masks, as shown in Fig. 4. Formally, the face mask attack on closed-set systems is formulated as the following optimization problem as a variation of Eq. (1) (formulations for other settings are presented in Appendix A):

$$\arg\max_{\boldsymbol{\delta}} \ell(S(\boldsymbol{x} + M \cdot \mathcal{T}(\boldsymbol{\delta})), y), \tag{3}$$

where $\boldsymbol{\delta} \in \mathbb{R}^{a \times b}$ is a $a \times b$ color matrix; each element of $\boldsymbol{\delta}$ represents an RGB color. M is the matrix that constrains the area of perturbations. $\mathcal{T}$ is a sequence of transformations that convert $\boldsymbol{\delta}$ to a face mask with a color grid in digital space by the following steps, as shown in Fig. 5. First, we use the *interpolation transform* to scale up the color matrix $\boldsymbol{\delta}$ into a color grid in a background image, which has the same dimension as $\boldsymbol{x}$ and all pixel values set to be 0. Then, we split the color grid into the left and right parts, each of which has four corner points. Afterward, we use a *perspective transformation* on each part of the grid for a 2-D alignment, which is based on the position of its source and destination corner points. Finally, we add the aligned color grid onto the input face image $\boldsymbol{x}$. Details of the perspective transformation and the algorithm for solving the optimization problem in Eq. (3) can be found in Appendix A.

3.2. Attacker's System Knowledge (K)

The key components of a face recognition system S are the training set D and neural architecture h. It is natural to ask how do these two components contribute to the robustness against adversarial attacks. From the attackers' perspective, we propose several evaluation scenarios in FACESEC,

which represent adversarial attacks performed under different knowledge levels on D and h.

Zero Knowledge. Both D and h are invisible to the attacker, *i.e.*, $K = \emptyset$. This is the weakest adversarial setting, as no critical information of S is leaked. Thus, it provides an upper bound for robustness evaluation on S. In this scenario, the attacks are referred to as *black-box attacks*, where the attacker needs no internal details of S to compromise it.

There are two general ways towards black-box attacks, *query-based attack* [5, 20] and *transfer-based attack* [21]. We employ the latter because the former attack requires a large number of online probes to repeatedly estimate the loss gradients of S on adversarial examples, which is less practical than fully offline attacks when access to prediction decisions is unavailable. The latter method is built upon the *transferability* of adversarial examples [21, 7]. Specifically, an attacker first collects a sufficient of training samples and builds a surrogate training set D'. Then, a surrogate system S' is constructed by training a surrogate neural architecture h' on D' for the same task as S, *i.e.*, $S' = f(h'; D')$. Afterward, the attacker obtains a set of adversarial examples by performing *white-box attacks* on the surrogate system S', which constitutes the transferable adversarial examples for evaluating the robustness of S.

Training Set. This scenario enables the assessment of the robustness of the training set of S in adversarial settings. Here, only the training set D is visible to the attacker, *i.e.*, $K = \{D\}$. Without knowing h, an attacker constructs a surrogate system S' by training a surrogate neural architecture h' on D, *i.e.,* $S' = f(h'; D)$. Then, the attacker performs the transfer-based attack aforementioned on S' and evaluates S by using the transferred adversarial examples.

Neural Architecture. Similarly, the attacker may only know the neural architecture h of S but has no access to the training set D, *i.e.*, $K = \{h\}$. This enables us to evaluate the robustness of the neural architecture h of S. Without knowing D, the attacker can build its surrogate system $S' = f(h; D')$ and conduct the transfer-based attack to evaluate S.

Full Knowledge. In the worst case, the attacker can have an accurate knowledge of both the training set D and neural architecture h (*i.e.*, $K = \{D, h\}$). Thus, it provides a lower bound for robustness evaluation on S. In this scenario, the attacker can fully reproduce S in an offline setting and then performs *white-box attacks* on S.

The evaluation method described above is based on the assumption that the adversarial examples in response to a surrogate system S' can always mislead the target system S. However, there is no theoretical guarantee, and recent studies show that some transferred adversarial examples can only fool the target system S with a low success rate [17].

To boost the transferability of adversarial examples produced on the surrogate system, we leverage two techniques: *momentum-based attack* [7] and *ensemble-based attack* [17, 7]. First, inspired by the momentum-based attack, we integrate the *momentum term* into the iterative process of the white-box attacks on the surrogate system S' to stabilize the update directions and avoid the local optima. Thus, the resulting adversarial examples are more transferable. Second, when the neural architecture h of the target system S is unavailable, we construct the surrogate system S' using an ensemble of models with different neural architectures rather than a single model, *i.e.*, $h' = \{h'_i\}_{i=1}^k$, where $\{h'_i\}_{i=1}^k$ is an ensemble of k models. Specifically, we aggregate the output logits of $h_i (i \leq k)$ in a similar way to [7]. The rationale behind this is that if an adversarial example can fool multiple models, it is more likely to mislead other models.

3.3. Attacker's Goal (G)

In addition to the attacker's system knowledge about S, adversarial attacks can differ in specific goals. In FACESEC, we are interested in the following two types of attacks with different goals:

Dodging/Non-targeted. In a dodging attack, an attacker aims to have his/her face misidentified as another arbitrary face. *e.g.*, the attacker can be a terrorist who wants to bypass a face recognition system for biometric security checking. As the dodging attack has no specific identity as which it aims to predict an input face image, it is also called the *non-targeted attack*.

Impersonation/Targeted. In an impersonation/targeted attack, an attacker seeks to produce an adversarial example that is misrecognized as a target identity. For example, the attacker may try to camouflage his/her face to be identified as an authorized user of a laptop, which uses face recognition for authentication.

In FACESEC, we formulate the dodging attack and impersonation attack as constrained optimization problems, corresponding to different face recognition systems and the attacker's goals, as shown in Table 1. Here, ℓ denotes the softmax cross-entropy loss used in closed-set systems, d represents the distance metric for open-set systems (*e.g.*, the cosine distance obtained by subtracting cosine similarity from one), $(\boldsymbol{x}, y)$ is the input face image and the associated identity, $\boldsymbol{\delta}$ is the adversarial perturbation, S represents a face recognition system which is built on either a single model or an ensemble of models with different neural architectures, M denotes the mask matrix that constrains the area of perturbation (similar to Eq. (1)), ϵ is the ℓ_p-norm bound of $\boldsymbol{\delta}$. For closed-set systems, we use y_t to represent the target identity of impersonation attacks. For open-set systems, we use $\boldsymbol{x}^*$ to denote the gallery face image that belongs to the identity as $\boldsymbol{x}$, and $\boldsymbol{x}_t^*$ as the gallery image for the target identity of impersonation.

Note that the formulations listed in Table 1 work for both digital attacks and physically realizable attacks: For the

Table 1. Optimization formulations by the attacker's goal.

Target System	Attacker's Goal	Formulation
Closed-set	Dodging	$\max_{\boldsymbol{\delta}} \ell(S(\boldsymbol{x}+M\boldsymbol{\delta}), y), \quad s.t. \ \|\|\boldsymbol{\delta}\|\|_p \leq \epsilon$
Closed-set	Impersonation	$\min_{\boldsymbol{\delta}} \ell(S(\boldsymbol{x}+M\boldsymbol{\delta}), y_t), \quad s.t. \ \|\|\boldsymbol{\delta}\|\|_p \leq \epsilon$
Open-set	Dodging	$\max_{\boldsymbol{\delta}} d(S(\boldsymbol{x}+M\boldsymbol{\delta}), S(\boldsymbol{x}^*)), \quad s.t. \ \|\|\boldsymbol{\delta}\|\|_p \leq \epsilon$
Open-set	Impersonation	$\min_{\boldsymbol{\delta}} d(S(\boldsymbol{x}+M\boldsymbol{\delta}), S(\boldsymbol{x}_t^*)), \quad s.t. \ \|\|\boldsymbol{\delta}\|\|_p \leq \epsilon$

former, we use a small value of ϵ and let M be an all-one matrix to ensure imperceptible perturbations on the entire image. For the latter, we use a large ϵ and let M to constrain $\boldsymbol{\delta}$ in a small area of $\boldsymbol{x}$.

3.4. Attacker's Capability (C)

In practice, even when the attackers share the same system knowledge and goal, their capabilities can still be different due to the time and/or budget constraints, such as the budget for printing adversarial eyeglass frames [26]. Thus, in FACESEC, we consider two types of attacks corresponding to different attacker's capabilities: *individual attack* and *universal attack*.

Individual Attack. The attacker has a strong capability with enough time and budget to produce a specific perturbation for each input face image. In this case, the optimization formulations are the same as those shown in Table 1.

Universal Attack. The attacker has a time/budget constraint such that he/she is only able to generate a *face-agnostic* perturbation that fools a face recognition system on a batch of face images instead of every input.

One common way to compute a universal perturbation is to sequentially find the *minimum* perturbation of each data point in the batch and then aggregate these perturbations [19]. However, this method requires orders of magnitude running time: it processes only one image at each iteration, so a large number of iterations are needed to obtain a satisfactory universal perturbation. Moreover, it only focuses on digital attacks and cannot be generalized to physically realizable attacks, which seek large perturbations in a restricted area rather than the minimum perturbations.

To address these issues, we formulate the universal attack as a *maxmin optimization* as follows (using the dodging attack on closed-set systems as an example):

$$\max_{\boldsymbol{\delta}} \min \{\ell(S(\boldsymbol{x}_i + M\boldsymbol{\delta}), y_i)\}_{i=1}^{N}, \quad s.t. \ ||\boldsymbol{\delta}||_p \leq \epsilon, \quad (4)$$

where $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$ is a batch of input images that share the universal perturbation $\boldsymbol{\delta}$. Compared to [19], our approach has several advantages: First, we can significantly improve the efficiency by processing images batchwise. Second, our formulation can explicitly control the universality of the perturbation by setting different values of N. Third, our method can be generalized to both digital attacks and physically realizable attacks. Details of our algorithm for solving the optimization problem in Eq. (4) and the formulations for other settings can be found in Appendix B.

Table 2. Open-set face recognition systems in our experiments.

Target Model	Training Set	Neural Architecture	Loss
VGGFace [23]	VGGFace [23]	VGGFace [23]	Triplet [23]
FaceNet [1]	CASIA-WebFace [37]	InceptionResNet [29]	Triplet [25]
ArcFace18 [2]	MS-Celeb-1M [10]	IResNet18 [14]	ArcFace [6]
ArcFace50 [2]	MS-Celeb-1M [10]	IResNet50 [14]	ArcFace [6]
ArcFace101 [2]	MS-Celeb-1M [10]	IResNet101 [14]	ArcFace [6]

4. Experiments

In this section, we evaluate a variety of face recognition systems using FACESEC on both closed-set and open-set tasks under different adversarial settings.

4.1. Experimental Setup

Datasets. For closed-set systems, we use a subset of the VGGFace2 dataset [3]. Specifically, we select 100 classes, each of which has 181 face images. For open-set systems, we employ the VGGFace2, MS-Celeb-1M [10], CASIA-WebFace [37] datasets for training surrogate models, and the LFW dataset [11] for testing.

Neural Architectures. The face recognition systems with five different neural networks are evaluated in our experiments: VGGFace [23], InceptionResNet [29], IResNet18 [14], IResNet50 [14], and IResNet101 [14].

Evaluation Metric. We use *attack success rate* = 1 - accuracy as the evaluation metric. Specifically, a higher attack success rate indicates that a face recognition system is more fragile in adversarial settings, while a lower rate shows higher robustness against adversarial attacks.

Implementation. For open-set face recognition, we directly applied five publicly available pre-trained face recognition models as the target models for attacks, as summarized in Table 2. At prediction stage, we used 100 photos randomly selected from frontal images in the LFW dataset [11], each of which is aligned by using MTCNN [38] and corresponds to one identity. And we used another 100 photos of the same identities as the test gallery. We computed the cosine similarity between the feature vectors of the test and gallery photos. If the score is above a threshold corresponding to a False Acceptance Rate of 0.001, then the test photo is predicted to have the same identity as the gallery photo.

For closed-set face recognition, we randomly split each class of the VGGFace2 subset into three parts: 150 for training, 30 for validation, and 1 for testing. To train closed-set models, we used standard transfer learning with the open-set models listed in Table 2. Specifically, we initialized each closed-set model with the corresponding open-set model, and then added a final fully connected layer, which contains 100 neurons. Unless otherwise specified, each model was

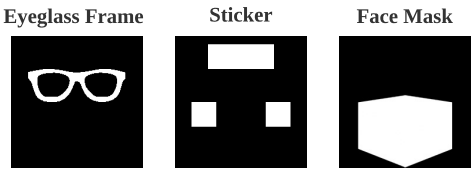

Figure 6. Mask matrices for physically realizable attacks in FACESEC.

trained for 60 epochs with a training batch size of 64. We used the Adam optimizer [13] with an initial learning rate of 0.0001, then dropped the learning rate by 0.1 at the 20th and 35th epochs.

For each physically realizable attack in FACESEC, we used $255/255$ as the ℓ_∞ norm bound for perturbations allowed, and ran each attack for 200 iterations. For the PGD attack [18], we used an ℓ_∞ bound $8/255$ and 40 iterations. The dimension of the color grid for face mask attacks is set to 16×8. The mask matrices that constrain the areas of perturbations for physically realizable attacks are visualized in Fig. 6.

4.2. Robustness of Face Recognition Components

We begin by using FACESEC to assess the robustness of face recognition components in various adversarial settings. For a given target face recognition system S and a perturbation type P, we evaluate the training set D and neural architecture h of S with the four evaluation scenarios presented in Section 3.2. Specifically, when h is invisible to the attacker, we construct the surrogate system S' by ensembling the models built on the other four neural architectures shown in Table 2. In the scenarios where the attacker has no access to D, we build the surrogate training set D' with another VGGFace2 subset that has the same classes as D in closed-set settings, and use the other four training sets listed in Table 2 for open-set tasks. We present the experimental results for dodging attacks on closed-set face recognition systems in Table 3, and the results for zero-knowledge dodging attacks on open-set VGGFace and FaceNet in Table 4. The other results can be found in Appendix C. Additionally, we evaluate the efficacy of using *momentum* and *ensemble* methods to improve transferability of adversarial examples, which is detailed in Appendix D.

It can be seen from Table 3 that: *the neural architecture is significantly more fragile than the training set in most adversarial settings*. For example, when only the neural architecture is exposed to the attacker, the sticker attack has a high success rate of 0.92 on FaceNet. In contrast, when the attacker only knows the training set, the attack success rate significantly drops to 0.01. In addition, by comparing each row of Table 3 that corresponds to the same target system, we observe that *digital attacks (PGD) are considerably more potent than their physically realizable counterparts on closed-set systems, while grid-level perturbations on face masks are noticeably more effective than pixel-level physically realizable perturbations (i.e., the eyeglass frame attack and the sticker attack).* Moreover, by comparing the zero knowledge attacks in Table 3 and 4, we find that *open-set face recognition systems are more vulnerable than closed-set systems* such that nearly all perturbation types of attacks (even the black-box sticker attack that often fails in closed-set) tend to be more likely to successfully transfer across different open-set systems (*i.e.*, these are more susceptible to black-box attacks), which should raise more concerns about their security.

Table 3. Attack success rate of dodging attacks on closed-set face recognition systems by the attacker's system knowledge. Z represents zero knowledge, T is training set, A is neural architecture, and F represents full knowledge.

Target System	Attack Type	Attacker's System Knowledge			
		Z	T	A	F
VGGFace	PGD	0.40	0.51	0.93	0.94
	Eyeglass Frame	0.23	0.28	0.70	0.99
	Sticker	0.05	0.06	0.47	0.98
	Face Mask	0.26	0.32	0.63	1.00
FaceNet	PGD	0.83	0.83	1.00	1.00
	Eyeglass Frame	0.13	0.16	0.90	1.00
	Sticker	0.01	0.01	0.92	1.00
	Face Mask	0.30	0.42	0.83	1.00
ArcFace18	PGD	0.87	0.92	0.97	1.00
	Eyeglass Frame	0.06	0.06	0.44	1.00
	Sticker	0.01	0.01	0.37	1.00
	Face Mask	0.27	0.33	0.71	1.00
ArcFace50	PGD	0.87	0.90	0.81	0.99
	Eyeglass Frame	0.09	0.12	0.44	0.99
	Sticker	0.00	0.01	0.14	0.94
	Face Mask	0.29	0.36	0.67	0.99
ArcFace101	PGD	0.81	0.78	0.86	0.96
	Eyeglass Frame	0.03	0.03	0.26	0.98
	Sticker	0.04	0.04	0.08	0.95
	Face Mask	0.26	0.36	0.54	0.99

Table 4. Attack success rate of dodging attacks on open-set face recognition systems with zero knowledge.

Target Model	Attack Type			
	PGD	Sticker	Eyeglass Frame	Face Mask
VGGFace	0.26	0.56	0.79	0.67
FaceNet	0.55	0.13	0.54	0.62

4.3. Robustness Under Universal Attacks

Next, we use FACESEC to evaluate the robustness of face recognition systems with various extents of adversarial universality by setting the parameter N in Eq. (4) to different values. For a given N, we split the testing set into mini-batches of size N, and produce a specific perturbation for each batch. Note that when $N = 1$, a universal attack is reduced to an individual attack. Table 5 shows the experimental results for universal dodging attacks on closed-set systems. The other results are presented in Appendix E.

Our first observation is that *face recognition systems are significantly more vulnerable to the universal face masks than other types of universal perturbations.* Under a large

Table 5. Attack success rate of dodging attacks on closed-set face recognition systems by the universality of adversarial examples. Here, N represents the batch size of face images that share a universal perturbation.

Target System	Attack Type	Attacker's Capability			
		N=1	N=5	N=10	N=20
VGGFace	PGD	0.94	0.86	0.31	0.15
	Eyeglass Frame	0.99	0.91	0.52	0.23
	Sticker	0.98	0.66	0.34	0.09
	Face Mask	1.00	1.00	0.88	0.56
FaceNet	PGD	1.00	1.00	0.80	0.21
	Eyeglass Frame	1.00	1.00	1.00	0.62
	Sticker	1.00	1.00	0.98	0.61
	Face Mask	1.00	1.00	1.00	0.91
ArcFace18	PGD	1.00	1.00	0.64	0.08
	Eyeglass Frame	1.00	0.96	0.44	0.08
	Sticker	1.00	0.56	0.09	0.00
	Face Mask	0.99	0.92	0.90	0.67
ArcFace50	PGD	1.00	0.80	0.37	0.05
	Eyeglass Frame	0.99	0.81	0.38	0.07
	Sticker	0.91	0.28	0.06	0.00
	Face Mask	0.99	0.98	0.81	0.72
ArcFace101	PGD	0.96	0.91	0.24	0.03
	Eyeglass Frame	0.98	0.71	0.19	0.02
	Sticker	0.93	0.15	0.03	0.00
	Face Mask	0.99	0.92	0.90	0.67

extent of universality (*e.g.*, when $N = 20$), face mask attacks remain > 0.5 success rates. Particularly noteworthy is the universal face mask attacks on FaceNet, which can achieve a rate as high as 0.91. In contrast, other universal attacks can have relatively low success rates (*e.g.*, 0.08 for eyeglass frame attack on ArcFace18). The second observation is that *the robustness of a face recognition system against different types of universal perturbations is highly dependent on its neural architecture*. For example, the ArcFace101 architecture is more robust than the others in most settings, while FaceNet tends to be the most fragile one.

4.4. Is "Robust" Face Recognition Really Robust?

While numerous approaches have been proposed for making deep neural networks more robust to adversarial examples, only a few [35] focus on defending against physically realizable attacks on face recognition systems. These defense approaches have achieved good performance for certain types of realizable attacks and neural architectures, but their effectiveness for other types of attacks and face recognition systems remains unknown. In this section, we apply FACESEC to evaluate the state-of-the-art defense paradigms. Specifically, we first use DOA [35], a method that defends closed-set VGGFace against eyeglass frame attacks [26] to retrain each closed-set system. We then evaluate the refined systems using the three physically realizable attacks included in FACESEC. Fig. 7 shows the experimental results for dodging attacks.

Our first observation is that *the state-of-the-art defense approach, DOA, fails to defend against the grid-level perturbations on face masks for most neural architectures*. Specifically, face mask attacks can achieve > 0.7 success rates on

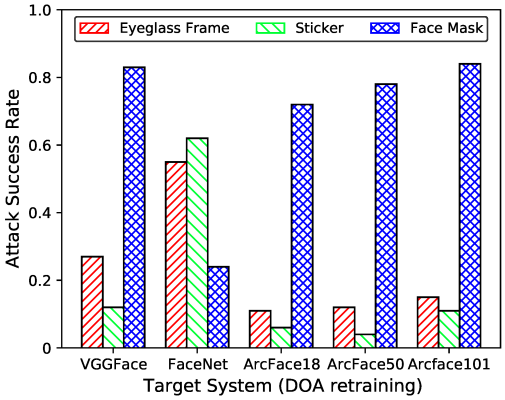

Figure 7. Attack success rate of dodging physically realizable attacks on closed-set systems with DOA retraining.

four out of the five face recognition systems refined by DOA. Moreover, we observe that *adversarial robustness against one type of perturbation can not be generalized to other types*. For example, while VGGface-DOA exhibits a relatively high level of robustness (more than a 70% accuracy) against pixel-level perturbations (*i.e.*, stickers and eyeglass frames), it is very vulnerable to grid-level perturbations (*i.e.*, face masks). In contrast, using DOA on FaceNet can successfully defend face mask perturbations with the attack success rate significantly dropping from 1.0 to 0.24, but it's considerably less effective against eyeglass frames and stickers. In summary, these results show that the effectiveness of defense is highly dependent on the nature of perturbation and neural architectures, which in turn, indicates that it is critical to consider different types of attacks and neural architectures when evaluating a defense method for face recognition systems.

5. Conclusion

We present FACESEC, a fine-grained robustness evaluation framework for face recognition systems. FACESEC incorporates four evaluation dimensions and can work on both face identification and verification of open-set and closed-set systems. To our best knowledge, FACESEC is the first-of-its-kind platform that supports to evaluate the risks of different components of face recognition systems from multiple dimensions and under various adversarial settings. The comprehensive and systematic evaluations on five state-of-the-art face recognition systems demonstrate that FACESEC can greatly help understand the robustness of the systems against both digital and physically realizable attacks. We envision that FACESEC can serve as a useful framework to advance future research of adversarial learning on face recognition.

Acknowledgement

Yevgeniy Vorobeychik was partially supported by the NSF (IIS-1905558 and ECCS-2020289) and ARO (W911NF1910241 and W911NF1810208).

References

[1] Facenet using pytorch. https://github.com/timesler/facenet-pytorch.

[2] Pytorch implementation of additive angular margin loss for deep face recognition. https://github.com/foamliu/InsightFace-v2.

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[4] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

[5] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.

[8] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019.

[9] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[10] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on Computer Vision*, pages 87–102. Springer, 2016.

[11] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[12] Anil K Jain and Stan Z Li. *Handbook of face recognition*, volume 1. Springer, 2011.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[14] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2811–2820, 2018.

[15] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

[16] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 507–516, 2016.

[17] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1773, 2017.

[20] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–169, 2018.

[21] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[22] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016.

[23] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12, September 2015.

[24] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing. *arXiv preprint arXiv:1906.07927*, 2019.

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[26] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael 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*, page 1528–1540, New York, NY, USA, 2016. Association for Computing Machinery.

[27] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014.

[28] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[29] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[31] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[32] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

[33] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[34] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[35] Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending against physically realizable attacks on image classification. In *8th International Conference on Learning Representations (ICLR)*, 2020.

[36] Xiao Yang, Dingcheng Yang, Yinpeng Dong, Wenjian Yu, Hang Su, and Jun Zhu. Delving into the adversarial robustness on face recognition. *arXiv preprint arXiv:2007.04118*, 2020.

[37] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.