Chapter 7
Adversarial Attacks on Face Recognition Systems

Ying Xu, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch

Abstract Face recognition has been widely used for identity verification both in supervised and unsupervised access control applications. The advancement in deep neural networks has opened up the possibility of scaling it to multiple applications. Despite the improvement in performance, deep network-based Face Recognition Systems (FRS) are not well prepared against adversarial attacks at the deployment level. The output performance of such FRS can be drastically impacted simply by changing the trained parameters, for instance, by changing the number of layers, subnetworks, loss and activation functions. This chapter will first demonstrate the impact on biometric performance using a publicly available face dataset. Further to this, this chapter will also present some strategies to defend against such attacks by incorporating defense mechanisms at the training level to mitigate the performance degradation. With the empirical evaluation of the deep FRS with and without a defense mechanism, we demonstrate the impact on biometric performance for the completeness of the chapter.

7.1 Introduction

Face recognition has been used in a large number of applications such as biometric authentication, civilian ID management and border crossing. The recent success of deep learning for recognition has led to very high biometric verification performance.

Y. Xu · K. Raja (✉) · R. Ramachandra · C. Busch
Norwegian University of Science and Technology, Trondheim, Norway
e-mail: kiran.raja@ntnu.no

Y. Xu
e-mail: xuyi@stud.ntnu.no

R. Ramachandra
e-mail: raghavendra.ramachandra@ntnu.no

C. Busch
e-mail: christoph.busch@ntnu.no

C. Rathgeb et al. (eds.), *Handbook of Digital Face Manipulation and Detection*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-87664-7_7

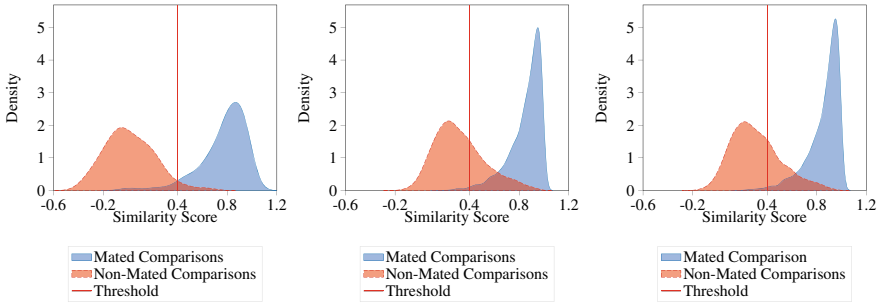
As a result, several state-of-the-art face recognition models such as VGGFace, Residual Networks (ResNet) and ArcFace have been extensively studied [2, 6]. The deeply learnt models have focused on improving the biometric performance in the presence of severe biometric sample quality degradation (i.e. face image) such as pose, illumination, expression, ageing and heterogeneity. With improved performance, the deep models can be used for identification where a subject is probed within the learnt models in a closed enrolment setting or for verification where the model is used to extract the features from two images and thereupon compare them to make a decision based on a pre-computed threshold.

In a parallel direction, a number of potential attacks have been reported on deeply learnt models for various tasks. The attacks range from simple perturbation in the input image to advanced attacks where the parameters of the model are changed. Such attacks lead to changing the robustness of the model; for instance, the changed input may lead to circumventing the identification (i.e. avoid identification from a black-list) or reaching a false match in a non-mated comparison trial. The attacks can be conducted in three different manners where an attacker is fully aware of the model's operation, partially aware of the model's operation and unaware of the model's operation, which fall under the categories white-box, black-box, and gray-box attacks [3, 4, 10, 11, 17, 21, 37, 40]. Each of these attacks can have different attack potential, and thus, not only making the deep models superior in terms of performance is needed, but demanding the robustness to be improved.

Several works have investigated the vulnerabilities of deeply learnt FRS for various attacks [1, 5, 8, 12, 16, 19, 22, 26–29, 32, 34, 35, 39, 42]. In this chapter, we provide a study on adversarial attacks on state-of-the-art deep Face Recognition System (FRS) based on ArcFace [6] in an open-set protocol setting, i.e. the testing set is unknown at the training level. We resort to such a protocol, given that most of the deeply trained FRS may be deployed in scenarios with unknown testing images. We provide a detailed analysis of the biometric implications when the attacks are successful, making the systems result in a higher False Match Rate (FMR). Specifically, when a threshold is set using a clean dataset for a fixed FMR, the attacks at the image level lead to higher FMR.

A sample illustration of such impact using two chosen attacks—Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) poisoning attacks is provided in Fig. 7.1 on a trained FRS using ArcFace [6]. As noted from the Fig. 7.1, a FRS working on the pre-defined threshold (in this case $\tau = 0.4$) for a fixed $FMR = 0.1\%$ will accept a score above such defined threshold in a non-mated comparison trial. The implication of such an attack is that an attacker can use a poisoned image to circumvent the verification process and thereby be verified as another subject. Such a case can be foreseen when a person contained in a watch-list can avoid being identified, putting the biometric FRS and, thereupon, the security at risk.

In order to fully illustrate the implications of such attacks, we employ FRGC v2 dataset to generate the attacks with FGSM and PGD. We limit the focus of the work to image level attacks under the assumption that the internals of the employed network is unknown to the attacker. To validate the attack potential, we consider the black-box attack setting on trained FRS models where the adversaries can attack



(a) Without Attacks (b) With FGSM Probe Attacks (c) With PGD Probe Attacks

Fig. 7.1 Illustration of increased False Match Rate due to fixed threshold based on the clean FRGC v2 Dataset

using perturbed/poisoned images[1] only at the testing/deploying stage. We use the clean version of the FRGC v2 dataset (i.e. with no poisoning) and the corresponding attack set to study the vulnerability. Further, we also re-train the model from scratch using the poisoned (attack) data as adversarial examples to make the trained model aware of such examples while learning. We further study the deep FRS models for their biometric performance with the trained models with adversarial examples. To provide an unbiased observation of the FRS, we employ disjoint training and testing sets without any subject overlap throughout the experiments in this chapter.

We conduct one study where an attacker has the full freedom to poison the probe data alone and another study where an attacker can also poison the enrolment data. In both cases, we assume that neither the trained model nor the training data set are available for the attacker to poison. Through empirical evaluations, we provide a detailed analysis and note the observations for the completeness of the chapter.

The main contributions of this chapter are

- Provides a detailed taxonomy of the potential adversarial attacks on the FRS and their applications.
- Provides empirical validation of vulnerability of the deeply learnt FRS model, which is trained from scratch. The attacks are generated through two different relevant and realizable approaches using Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD).
- Provides a comparative evaluation of the deeply learnt FRS model against commercial-off-the-shelf (COTS) FRS to benchmark the impact of the adversarial attack in each case.
- Provides an evaluation of the robustness of FR models when the same is trained with the adversarial examples using FGSM and PGD.

[1] Both perturbed and poisoned images in this chapter refer to the same kind of attacks and are used interchangeably.

In the rest of this chapter, we first list out the taxonomy of the potential adversarial attacks on FRS in Sect. 7.2 and provide the details on the chosen attacks for the evaluation in Sect. 7.3. We then provide the details of the deeply learnt FRS in Sect. 7.5 followed by the details on empirical evaluation in Sect. 7.6. We provide the discussion on the observations in Sect. 7.8 and conclude the chapter with potential research directions.

7.2 Taxonomy of Attacks on FRS

Szegedy et al. [33] illustrated the impact of small perturbations on the images for the image classification problem and defeated state-of-the-art Deep Neural Networks (DNNs) with high misclassification rates. These misclassified samples were named adversarial examples that can impact the performance of the deep models. A number of works have thereafter been proposed for creating such attacks, and the adversarial attacks can be classified by the amount of knowledge an attacker has over the model [3, 4, 10, 11, 17, 21]. Based on such knowledge, the attacks can be classified [37, 40] as:

- White-box attack—assuming the complete knowledge of the target model, i.e. its parameters, architecture, training method, and even in some cases, its training data.
- Gray-box attacks—having partial knowledge of the internal operations and parameters of the network.
- Black-box attacks—feeding a target model with the adversarial examples (during testing) created without knowing that model (e.g. its training procedure or its architecture or parameters). Despite the limited knowledge of the model, an attacker can interact with such a model by utilizing the transferability of adversarial examples.

Motivated by such adversarial attacks, several works have investigated the impact of such attacks on FRS and have provided various mitigation measures [1, 5, 8, 12, 16, 19, 22, 27–29, 32, 34, 35, 39, 42]. We provide an alternative taxonomy of such adversarial attacks by categorizing them in two dimensions such as threat model and perturbation. Figure 7.2 presents the taxonomy under two such dimensions with various sub-attacks. We provide a brief overview of the attacks for the convenience of the reader in this section.

7.2.1 Threat Model

We could break down the threat model into four perspectives, adversarial falsification, adversary's knowledge, adversarial specificity and attack frequency, making different attack examples from various kinds of adversarial attack attributes ground on different assumptions, the knowledge of the model, specificity and attack scenarios.

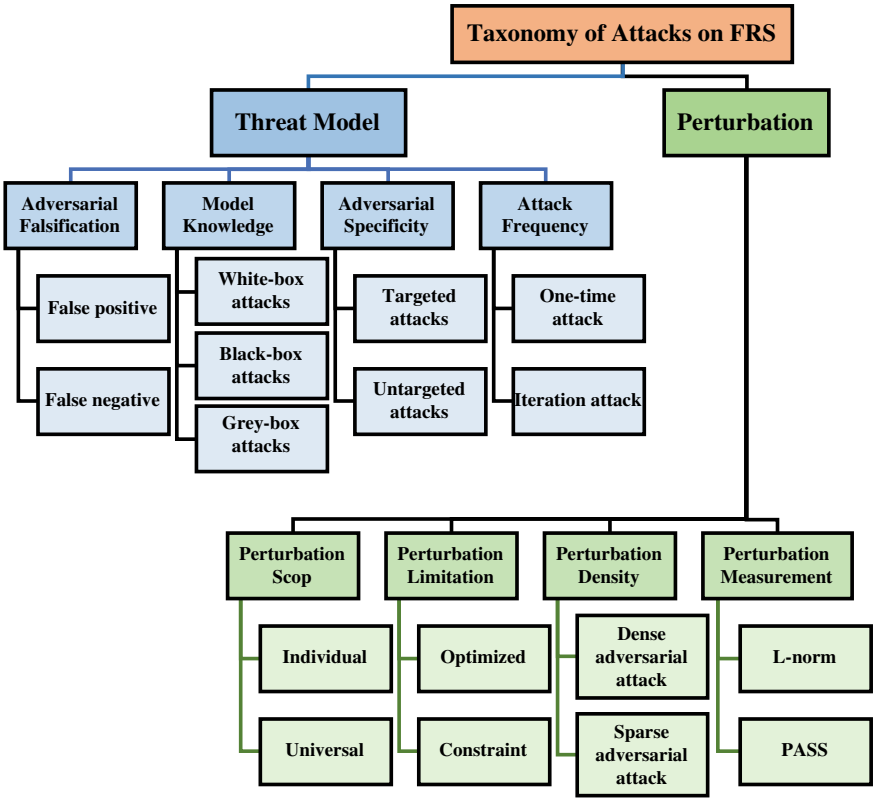


Fig. 7.2 A taxonomy of potential adversarial attacks on FRS

(A) Adversarial Falsification

(i) **False positive**: A false positive attack rejects a true null hypothesis, also called Type I Error, where a negative example is misclassified as a positive class. Within the context of FRS, this error implies a comparison decision of match for a biometric probe and a biometric reference from different biometric capture subjects. For instance, a false match is when subject A is identified or falsely verified as subject B, i.e. a zero-effort impostor accepted in a non-mated comparison trial.

(ii) **False negative**: A false negative attack makes the non-rejection of a false null hypothesis, also called Type II Error, where a positive example is misclassified as a negative class. In the context of FRS, this implies a comparison decision of "non-match" for a biometric probe and a biometric reference from the same biometric capture subject and the same biometric characteristic. Alternatively, a subject A in a mated comparison trial is rejected by the biometric system.

(B) Model Knowledge

(i) **White-box attacks**: A white-box attack gets all the information and parameters, including the model architectures, model weights, activation functions and all other hyper-parameters inside the machine learning model to attack, and generates adversarial samples based on the gradient of the given model.

(ii) **Black-box attacks**: A black-box attack generates adversarial samples only by the knowledge of the inputs and the outputs of a neural network model. For example, when an adversarial image is provided to the model, a label or a confidence score corresponding to another class of image is returned based on the chosen model. Black-box attacks can be divided into transfer-based, score-based and decision-based attacks. An evolutionary attack method for query-efficient adversarial attacks in the decision-based black-box setting [7] is proposed to optimize attack objective function in a black-box manner through queries only.

(iii) **Grey-box attacks**: A grey-box attack is an intermediate attack that lies between former and latter attacks. Typically in grey-box attacks, an attacker can exploit partial knowledge of models, inputs and outputs of a neural network model.

(C) Adversarial Specificity

(i) **Targeted attacks**: The targeted attack changes the output classification of input to the desired one. For example, many different attacks can be conducted to be verified or identified as another subject. Dodging attacks is such kind of attacks where the face can be accessorized with glasses or makeup to be identified as another subject [32].

(ii) **Untargeted attacks**: The goal of an untargeted attack is to lead the neural network to misclassify the inputs. An attacker can simply employ similar approaches of wearing a mask, glasses [32], makeup [42] or have expressions [22] to impersonate another subject, typically an enrollee within the enrolment dataset.

(D) Attack Frequency

(i) **One-time attack**: A one-time attack takes only one time to raise the adversarial examples. A number of different approaches can be used for circumventing the FRS, for instance, creating a face image through deepfakes [16, 27, 34].

(ii) **Iteration attack**: An iterative attack takes multiple times to upgrade the adversarial examples. A potential use case of such attacks can be in creating a morphed face image by combining two face images iteratively with various morphing factors until a successful verification is obtained [25, 36].

Perturbation

Adding perturbations on face images is an easy but effective attack on FRS. Adversarial examples could be generated by adding a small imperceptible perturbation to deceive both humans and the model. Although larger perturbations can be added to

the face images, this will lead to producing non-human figures, and the applicability of such perceptible perturbations can only fool FRS but not the human operators if such a system is monitored by one. The perturbation could be categorized in three different sets based on factors of perturbation scope, perturbation limitation, and perturbation measurement.

(A) Perturbation Scope

(i) **Individual perturbation**: Individual attacks produce various perturbations for each clean input. For instance, a face image may be blurred, added pixel-level noises, masked portions of the face to create the adversarial sample [1, 12].

(ii) **Universal perturbation**: Universal attacks generate a universal perturbation for the entire data set. Although these attacks are very effective, an attacker needs to avail the entire dataset to devise a good perturbation model to fool the FRS effectively [1, 12, 19, 41].

(B) Perturbation Limitation

(i) **Optimized perturbation**: An optimized perturbation aims to minimize the perturbation in order to prevent humans from recognizing the perturbation, in the meantime, to fool the FRS [29, 39].

(ii) **Constraint perturbation**: A constraint perturbation, on the other hand, sets perturbation as a diminutive constraint, for instance, in a chosen area of the face [5, 24].

(C) Perturbation Density

(i) **Dense adversarial attack**: Dense adversarial attacks perturb the image over all the pixels in one image [3]. As the perturbations are spread over the image, these attacks can be effective, but when the perturbation level is increased, the image structure may change, making them irrelevant attack samples mainly due to loss of visual fidelity.

(ii) **Sparse adversarial attack**: A sparse adversarial attack means only partial positions are considered, regardless of those immaterial pixels. The adversarial model would choose which parts should be attacked. Perturbation factorization [8] was proposed to enable sparse, dense adversarial attacks.

(D) Perturbation Measurement

(i) **ℓ_p-norm:** ℓ_p-norm is used to define the magnitude of perturbations which is denoted as $\|\mathbf{x}\|_p$ on a vector $\mathbf{x}$ and is defined as

$$\|\mathbf{x}\|_p = \sqrt[p]{\sum_{n}^{i=1} |v_i|^p} \tag{7.1}$$

where p defines the norm. The one-norm (also known as the L_1-norm, ℓ_1-norm, or mean norm), where p equals 1, is defined as the sum of the

absolute values of its components. The two-norm (also known as the L_2-norm, ℓ_2-norm, least-squares norm or mean norm), where p equals 2, is defined as the square root of the sum of the squares of the absolute values of its components. The infinity norm (also known as the L_∞-norm, ℓ_∞-norm, max norm, or uniform norm), where p equals ∞, is defined as the maximum of the absolute values of its components.

(ii) **Psychometric perceptual adversarial similarity score (PASS)**: A novel Perceptual Adversarial Similarity Score (PASS) [28] is a new measure to quantify adversarial images. It is proposed to be more consistent with human perception than prior ℓ_p-norm measurements and to serve as a similarity measure to quantify how adversarial a misclassified image is. It supports many transformations, including small translations and rotations, which result in images that are perturbed to observable extents compared to their original counterparts while still appear to be reasonable samples of the same images.

7.3 Poisoning Attacks on FRS

Although several attacks can be found in the literature, we focus on "Adversarial Falsification" attacks under which both False Non-Match Rate (FNMR) and False Match Rate (FMR) are impacted. Further, we restrict ourselves to Black-box setting where the knowledge of the model is limited and create the attacks using perturbations (or poisoning). Two kinds of perturbations such as Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) [11] attacks are considered in this chapter, mainly due to lower attack generation cost in terms of time and effort. Different variants of the same attacks can be found in the literature, but they generally take a longer time to generate, and we restrict our focus to realizable attacks in terms of the time required to generate the attack itself. We provide a brief overview of the attack generation mechanism for both attacks in this section.

7.3.1 Fast Gradient Sign Method

The Fast Gradient Sign Method (FGSM) [11] is a linear perturbation of non-linear models. It uses the gradients of the neural network to create adversarial examples. The perturbation is defined as

$$\eta = \epsilon sign\left(\nabla_x J\left(\theta, x, y\right)\right),$$

where θ is the parameters of a model, x and y are the input to the model and the labels associated with x respectively, $J\left(\theta, x, y\right)$ represents the cost used to train the neural network and ϵ is the perturbation factor. The optimal max-norm η is defined by

linearizing the cost function around the current value of θ. The adversarial image is produced by adding η to the original input image. The neural networks are designed by leveraging the gradients to optimize the learning. The FGSM attack generation simply uses the gradient of loss of the input data and adjusts the input data in such a way that the loss is maximized.

7.3.2 Projected Gradient Descent

The idea of Projected Gradient Descent (PGD) [11] is essentially a saddle point problem as the composition of an inner maximization problem and an outer minimization problem. The basic formulation of PGD is denoted as

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}(\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y)).$$

$\mathcal{D}$ represents an underlying data distribution over pairs of examples x and corresponding labels y. The θ is the set of model parameters and $L(\theta, x, y)$ is the loss function. The goal of PGD algorithm is to find parameters θ that minimize the empirical risk $\mathbb{E}_{(x,y)\sim\mathcal{D}}(L(\theta, x, y))$. A set of allowed perturbations $\mathcal{S}$ is introduced to formalize the manipulative power of the adversary for each data point x. $\mathcal{S}$ captures perceptual similarity between images in the classification tasks. The goal of the inner maximization problem is to find a perturbation $\delta \in \mathcal{S}$ of a given data point x that achieves the highest loss. While the outer minimization problem aims to find the model parameters to minimize the adversarial loss. PGD algorithm can start from random perturbations in the ball of interest decided by ℓ_∞-norm around a sample and repeatedly take s steps of α size till convergence. Random starts would help PGD to solve local optima within the objective.

7.4 Carlini and Wagner (CW) Attacks

The general idea of CW algorithm [3] is the typical adversarial attack which utilizes the adversarial loss and the image distance loss. The former loss ensures the adversarial images to fool the classification models while the latter one is used to control the perturbation of the adversarial examples. The CW attack could be formulated as

$$\text{minimize} \quad \|\delta\|_p + c \cdot f\,(x+\delta) \tag{7.2}$$

$$\text{such that} \quad x+\delta \in [0,1]^n \tag{7.3}$$

c is a constant that differs between models. The author of CW used binary search to choose c. δ is the small change that the CW algorithm adds to mystifies the classifier. Given x_i, δ_i is defined as

$$\delta_i = \frac{1}{2}(tanh\,(w_i) + 1) - x_i,$$

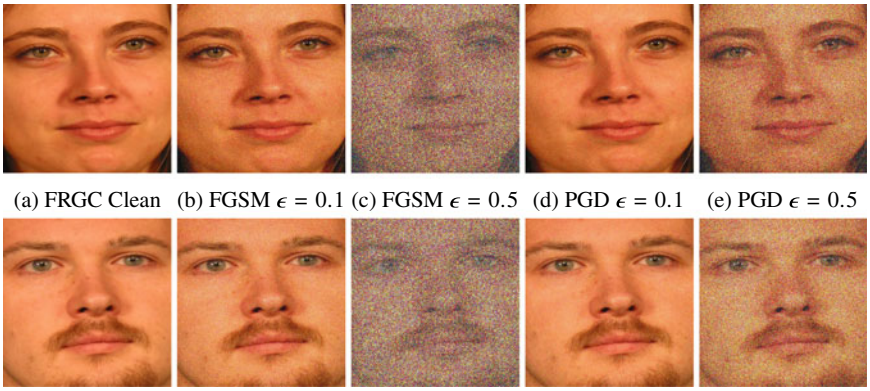
$tanh(w_i)$ is introduced to meet the request of box constraint Eq. (7.3).

Object function f is chosen as

$$f(x') = \max(\max_{i \neq t} Z(x')_i - Z(x')_t, -k),$$

which chooses the difference of two probability values or the confidence parameter k. By setting the value of $-k$, the user could specify the confidence of the adversarial attack. This chapter focuses on open-set verification protocols by simply extracting the embeddings and comparing with cosine distance, and therefore we do not consider this attack further.

A sample illustration of FGSM and PGD perturbation is shown in Fig. 7.3. As noted from Fig. 7.3, perturbation factor ϵ directly influences the perceptual quality of the image. While higher perturbation factors may result in stronger attacks, one has to focus on visual appearance to make the attack not obvious to human perception.



(a) FRGC Clean (b) FGSM $\epsilon = 0.1$ (c) FGSM $\epsilon = 0.5$ (d) PGD $\epsilon = 0.1$ (e) PGD $\epsilon = 0.5$

(f) FRGC Clean (g) FGSM $\epsilon = 0.1$ (h) FGSM $\epsilon = 0.5$ (i) PGD $\epsilon = 0.1$ (j) PGD $\epsilon = 0.5$

Fig. 7.3 Adversarial attack examples of FGSM and PGD with $\epsilon = 0.1$ and $\epsilon = 0.5$ where ϵ is the strength of the perturbation. As noted from the illustration, FGSM attack degrades the visual appearance quality of the image when the perturbation factor ϵ is increased while the visual appearance is still tolerable with the PGD even with a larger ϵ

7.5 ArcFace FRS Model

Of the number of models available for large-scale training data, both the softmax-loss-based methods [2] and the triplet-loss-based methods [30] can achieve high recognition performance. However, both the softmax loss and the triplet loss have some drawbacks for scalability issues. The size of the linear transformation matrix $W \in \mathbb{R}^{d \times n}$ increases linearly with the identities number n, and the learned features are separable for the closed-set classification problem but not discriminative enough for the open-set face recognition problem which is typical for face recognition. As for the triplet loss, the combinatorial explosion in the number of face triplets is especially for large-scale datasets, leading to a significant increase in the number of iteration steps. Semi-hard sample mining is a quite difficult problem for effective model training, which depends on the availability of large-scale data. Based on these two motivations, we choose to employ ArcFace deep FRS due to its superior performance as demonstrated in various works [6].

In this work, we choose to employ the ResNet101 architecture and Additive Angular Margin Loss (ArcFace) loss to directly benefit from the discriminative power of the face recognition model without much overhead on training process [6]. ArcFace utilizes the arc-cosine function to calculate the angle between the current feature and the target weight. ArcFace directly optimizes the geodesic distance margin under the exact correspondence between the angle and arc in the normalized hypersphere. Specifically, we extract 512 dimensional embeddings for all the experiments.

We first validate the choice of ResNet101 network and ArcFace loss using the publicly available LFW [13], CFP-FP [31], AgeDB-30 [20]. Based on the accuracy obtained on these datasets, we fix the architecture choices and then use it for all our experiments on FRGC v2 dataset [23].

7.6 Experiments and Analysis

In this section, we list the details of the dataset, attack generation and the set of FRS analysis conducted. We employ False Non-Match Rate (FNMR) at a False Match Rate (FMR) of 0.1% and Equal Error Rate (EER) to report the performance of FRS and supplement the results using the Detection Error Trade-off (DET) curves when applicable.

7.6.1 Clean Dataset

Considering the focus of this work on FRS, we choose a state-of-art FR dataset–FRGCv2 dataset [23] specifically to report the open-set verification experiments. Our choice is based on two factors (1) FRGCv2 dataset presents a mix of images that

closely resemble the biometric enrolment and probe dataset and are not significantly degraded, impacting the model's performance due to noise (2) by splitting the FRGC dataset into disjoint sets, we can illustrate the performance on open-set verification protocols. We, therefore, evaluate the attack potential on the deeply learnt FRS model corresponding to the protocol known as Experiment-1 [23]. We have reorganized the dataset to have 222 subjects in the training set and validation set (randomly subsampled in each training epoch) and the rest of the non-overlapping subjects in the disjoint testing test. Care has been exercised not to overlap any subjects in the training set and testing set. The database is first processed to detect the face region, and then the facial images are aligned [6]. Each image in all three sets is further resized to 112×112 pixels for training the model and testing the model.

7.6.2 Attack Dataset

We generate the attack dataset corresponding to all three subsets, such as training, validation and testing set of FRGC v2 dataset. We generate two kinds of attacks such as FGSM attacks and PGD attacks as both of these attacks can retain the similarity of the face region despite adding the noise to the image.[2]

7.6.2.1 Attack Dataset—FGSM Perturbations

Using the clean version of the FRGC dataset (i.e. non-poisoned), we generate the FGSM attack dataset for all three subsets of training, validation and testing set. We employ Torchattack library[3] to generate the attack dataset for FGSM. We specifically use the FGSM model from Torchattack library to generate the attacks with a perturbation factor of $\epsilon = 0.1$ and $\epsilon = 0.5$.[4] Although we have experimented with various ϵ, we choose the perturbation factor of $\epsilon = 0.1$ based on the stronger attack potential while not degrading the image's visual appearance. It should, however, be noted that the $\epsilon < 0.1$ is still effective to attack FRS with a limited success rate.

7.6.2.2 Attack Dataset—PGD Perturbations

Similar to FGSM attacks, we use the clean version (i.e. non-poisoned) of the FRGC dataset to generate a PGD attack dataset for all three subsets of training, validation and testing set. We employ the Torchattack library to generate the PGD attack dataset.

[2] CW attacks take larger time for generation, and we do not consider CW attacks in this work as their practical applicability in our study is limited.

[3] https://adversarial-attacks-pytorch.readthedocs.io/en/latest/attacks.html.

[4] https://github.com/Harry24k/adversarial-attacks-pytorch.

In the lines of FGSM attacks, we employ generate the attacks with a perturbation factor of $\epsilon = 0.1$ and $\epsilon = 0.5$.

7.6.2.3 COTS Evaluation

In order to first understand the impact of poisoning (perturbation) attacks, we evaluate the biometric performance using the COTS system.[5] We employ testing partition of clean FRGC data and testing partition of poisoned data with FGSM and PGD attacks to verify the recognition performance. We first evaluate the performance of COTS FRS using clean FRGC data against clean data. We further evaluate the performance of COTS FRS by enrolling clean FRGC data and probed using PGD and FGSM attacks generated with $\epsilon = 0.1$. The attacks generated with $\epsilon = 0.5$ do not compromise the FRS as the FRS rejects them as Failure-to-Extract, and we do not report the error rates for such a setting.

We note from the Table 7.1 that COTS FRS[6] is not sensitive to the poisoned data and provides ideal biometric performance irrespective of clean or poisoned data. Our assertion of this observation is that the version of the COTS FRS does not employ deep networks and thus makes it robust against poisoning attacks. However, as the COTS FRS does not disclose the algorithm, we cannot fully confirm our hypothesis.

7.6.3 FRS Model for Baseline Verification

We train the ArcFace deep learning model from scratch using the training set and verify the model's performance using the disjoint validation set. We carry out the training for 100 epochs with a learning rate of 0.01 with ArcFace loss [6] to avoid overfitting due to limited sample size. The trained model is further used to extract the embedding of length 512 on the testing set, and the similarity between two images is computed using the cosine distance in our baseline performance evaluation. We employ the False Non-Match Rate (FNMR) at False Match Rate (FMR) of 0.01 for validating the model on the validation set. The performance reported in this chapter further on is only on the testing set of the FRGCv2 dataset and corresponding attack sets for FGSM and PGD attacks.

7.6.4 FRS Baseline Performance Evaluation

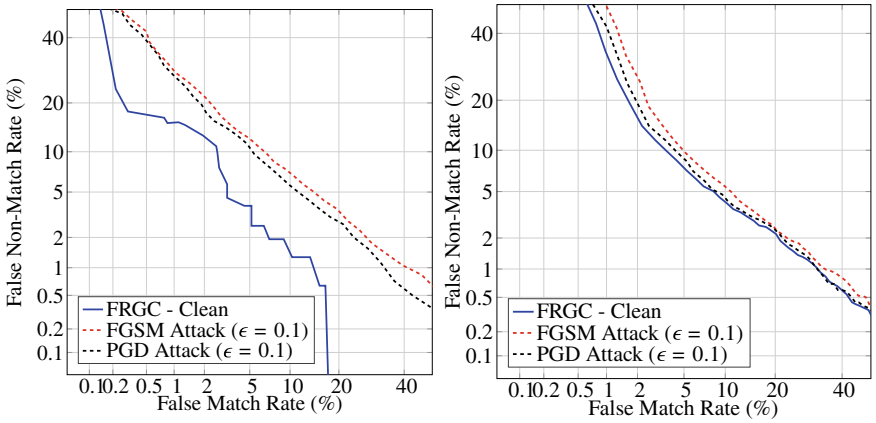
The trained model on FRGC v2 training dataset is first evaluated to obtain the baseline performance on the FRGC testing set, FGSM attack testing set and PGD attack

[5] Neurotech Verilook—Version 11.1—https://www.neurotechnology.com.

[6] We do not present the DET curves as the EER=0% for chosen COTS SDK.

Table 7.1 Performance of FRS without attacks, with FGSM attacks and PGD attacks

	EER (%)	FNMR (%) @ FMR = 0.1%	EER (%)	FNMR (%) @ FMR = 0.1%
Deep FRS	Cosine Similarity		Euclidean Distance	
FRGC Clean	4.18	13.54	6.21	42.68
FGSM probe attacks	8.08	39.44	6.59	74.04
PGD probe attacks	7.45	31.98	7.20	84.32
COTS FRS				
	EER (%)		FNMR (%)	
FRGC Clean	0		0	
FGSM probe attacks	0		0	
PGD probe attacks	0		0	



(a) DET curves - baseline performance for models trained on FRGC clean data (Cosine similarity)

(b) DET curves - baseline performance for models trained on FRGC clean data (Euclidean distance)

Fig. 7.4 Baseline DETs on FRS trained on clean FRGC, probed with FRGC clean data, FGSM and PGD attack data

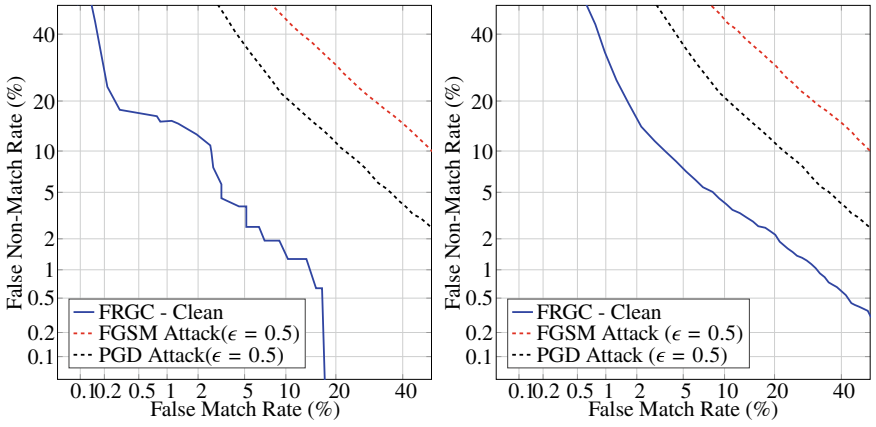
testing set. The results obtained from baseline evaluation are presented in Fig. 7.4a. For reporting the performance, we extract the embedding of length 512 from the trained FRS and then employ cosine similarity to obtain the comparison score. As noted from Table 7.1 and the corresponding DET can be found in Fig. 7.4a, the trained model performs best when the data is clean (i.e. without attack), resulting in an Equal Error Rate (EER) of 4.18%.

7.6.4.1 Baseline Evaluation with Euclidean Distance

In order to study the variance of performance with distance measure on the FRS model, we also conduct the same analysis using the Euclidean distance measure to obtain the comparison scores. As it can be observed from Fig. 7.4b, there is a performance drop when the embeddings are compared using the Euclidean distance illustrating the dependence of distance measure in deep FRS. This aspect can be attributed to the training mechanism optimized for cosine similarity, and thus it is not surprising to see the drop in the performance. Table 7.1 presents the obtained error rates using the Euclidean distance with a baseline EER of 6.21% when the model is presented with no attacks.

7.6.4.2 Impact of Increased Perturbations

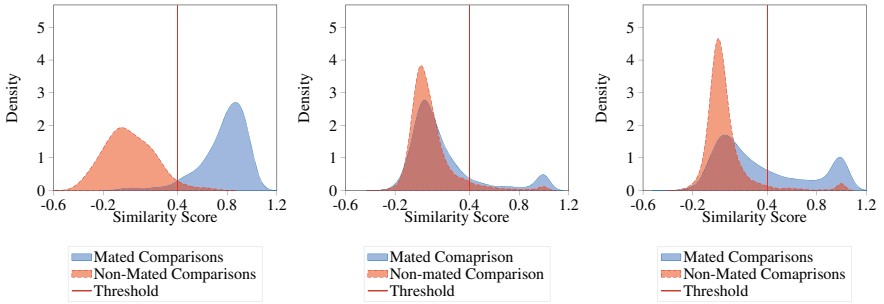
Further, we also study the impact of the perturbation strength on FRS by poisoning the images with a perturbation factor ϵ of 0.5. Specifically, we poison the probe images and use them to probe against clean FRGC enrolment. Figure 7.5a presets the DETs corresponding to these experiments and it can be noted from the Fig. 7.5a that such attacks lead to a significant number of false matches and false non-matches. A similar observation can be made for the comparison of embeddings using the Euclidean distance as depicted in Fig. 7.5b. Further, to illustrate the impact of such attacks with a high degree of poisoning, we present the distribution shifts Fig. 7.6. As one can note, such attacks lead to very high false rejects and a small number



(a) DET curves - baseline performance for models trained on FRGC clean data (Cosine similarity) - Attacks FGSM (ϵ = 0.5) and PGD (ϵ = 0.5)

(b) DET curves - baseline performance for models trained on FRGC clean data (Euclidean distance) - Attacks FGSM (ϵ = 0.5) and PGD (ϵ = 0.5)

Fig. 7.5 Baseline DETs on FRS trained on clean FRGC and tested on FRGC clean data, FGSM and PGD attack data with $\epsilon = 0.5$



(a) Without Attacks (b) With FGSM Probe Attacks (c) With PGD Probe Attacks

Fig. 7.6 Illustration of increased False Non-Match Rate due to fixed threshold based on the clean FRGC v2 Dataset and probed with highly perturbed images $\epsilon = 0.5$

Table 7.2 Performance of FRS with FGSM and PGD attacks with larger perturbation ($\epsilon = 0.5$)

	EER (%)	FNMR (%) @ FMR = 0.1%	EER (%)	FNMR (%) @ FMR = 0.1%
	Cosine		Euclidean	
FGSM probe attacks	24.72	86.20	18.44	89.28
PGD probe attacks	15.05	65.46	24.81	74.50

of false matches. Table 7.2 presets the performance obtained in terms of EER and FNMR@FMR=0.1% to illustrate the degradation of FRS.

The attacks with such amount of poisoning may not benefit the attacker to be falsely verified against another identity, making them not highly lucrative for the attackers targeting false acceptance. However, such attempts for verification using highly poisoned images may easily help the attacker to be not identified in a watch-list where the FRS does not obtain a high enough comparison score to cross the pre-determined threshold. It can be asserted with a high degree of confidence that this kind of attacks may not be attractive as they distort the images to a high degree.

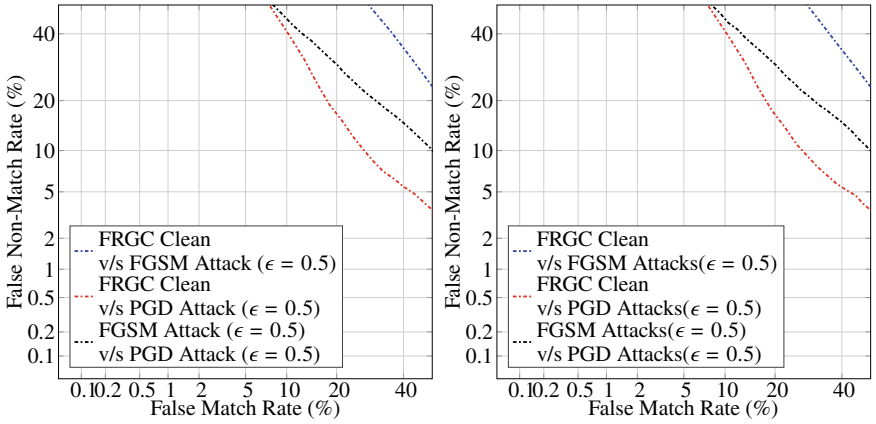
7.6.5 FRS Performance on Probe Data Poisoning

Considering that an attacker is unable to change the enrolment set, we also provide another study where the attacker can only change the data at the probe level. The critical assumption here is that an attacker can get hold of images from social media sites that may not be of optimal quality for biometric use cases. Using such images, an attacker can generate the poisoning such that the FRS can still accept the attack images. In order to achieve this, we retain the original FRGC clean data as an enrol-

ment set and use the FGSM and PGD attacks at the probe level. Figure 7.4 presents the change in performance when the probe images are alone attacked where the poisoned data succeeds in verifying against the enrolment set. This can be both seen as the robustness of the network to noisy data and also as a weakness in distinguishing the poisoned attack images.

7.6.6 FRS Performance on Enrolment Data Poisoning

While we have assumed that an attacker is unable to access the enrolment set in the earlier set of experiments, we also consider another scenario where the attacker is fully capable of poisoning the enrolment dataset. We consider a scenario where an attacker can poison the enrolment database using FGSM attacks and probe against PGD attacks. As illustrated in Fig. 7.7a and b, under such a scenario of poisoned enrolment set, the attack succeeds in obtaining a reasonable biometric performance. However, these attacks may not be highly realistic when secure mechanisms are used to protect the enrolment data, as seen in most of the operational systems. Despite limited success, this set of experiments shows that the FRS are vulnerable if the enrolment set is compromised, and this aspect needs further investigation.



(a) Cross-poisoning evaluation of FRGC clean data v/s FGSM ($\epsilon = 0.5$) and PGD ($\epsilon = 0.5$) attack - Cosine Distance

(b) Cross-poisoning evaluation of FRGC clean data v/s FGSM ($\epsilon = 0.5$) and PGD ($\epsilon = 0.5$) attack - Eucliedean Distance

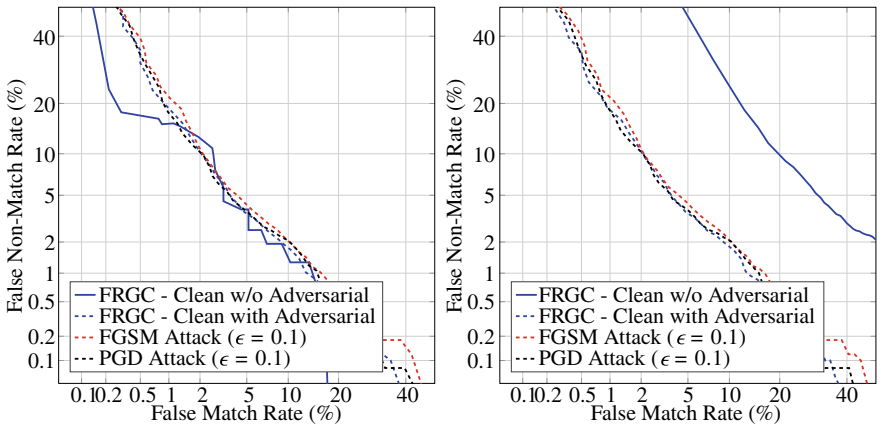
Fig. 7.7 DETs for clean FRGC enrolment poisoned with versus attack probe images with higher perturbations ($\epsilon = 0.5$)

7.7 Impact of Adversarial Training with FGSM Attacks

As the performance of the FRS under adversarial attacks can change, in this section, we analyze if training the FRS with adversarial samples can improve the accuracy. While different strategies for mitigating the adversarial attacks starting from having detection schemes [18] to training the FRS with adversarial samples [9, 38], we simply resort to train the FRS model with the adversarial samples using both perturbation factors of $\epsilon = 0.1$ and $\epsilon = 0.5$. To account for the generalisability towards both FGSM and PGD attacks, we train a FRS network by incorporating the FGSM and PGD adversarial samples into the training data.

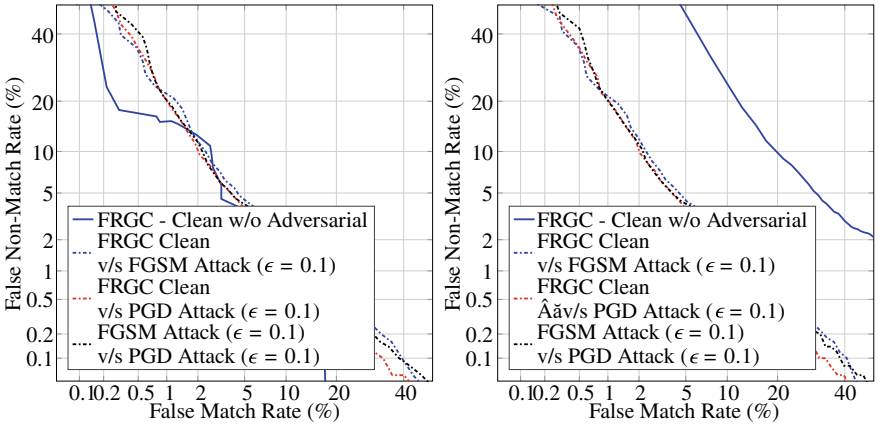
Figure 7.8 depicts the performance obtained using the FRS trained with FGSM + PGD attacks on the various testing sets. As it can be noted, the FRS, despite having low accuracy when the adversarial samples are presented under open-set evaluation protocol, performance is restored to similar accuracy simply by incorporating the adversarial samples in the training set. It is interesting to note that the adversarially trained model performs equally well with the embeddings compared using Euclidean distance, unlike the model trained with clean data under similar settings as shown in the Fig. 7.9. Although this indicates the robustness of the trained model when adversarial samples are provided, a detailed analysis is further needed.

Further, we also evaluate the performance of the adversarially trained FRS for cross-poisoning attacks corresponding to Sect. 7.6.5. The obtained performance is presented in the Fig. 7.9 and the performance is also listed in Table 7.3. It can be evidently noted that adversarial training can help in addressing the cross-poisoning attacks to a greater extent. In the lines of previously noted results, it can be seen that the adversarial training also improves the performance for comparison scores obtained with the Euclidean distance measure for measuring the dissimilarity between embed-



(a) FGSM+PGD Trained - Cosine Similarity (b) FGSM+PGD Trained - Euclidean Distance

Fig. 7.8 ROC graphs for adversarial trained FRS with FGSM+PGD attack data



(a) FGSM+PGD Trained - Cosine Similarity (b) FGSM+PGD Trained - Euclidean Distance

Fig. 7.9 ROC graphs for adversarial trained FRS with FGSM+PGD attack data and increased poisoning

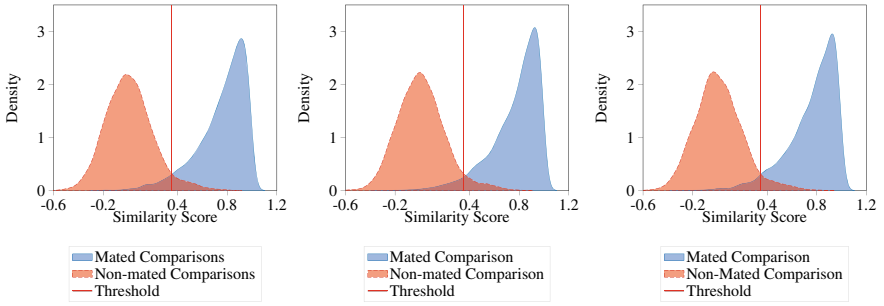
Table 7.3 Performance of FRS trained with adversarial examples when probed with attack images from FGSM and PGD attack generation

	EER (%)	FNMR (%) @ FMR = 0.1%	EER (%)	FNMR (%) @ FMR = 0.1%
	Cosine		Euclidean	
FRGC Clean	4.18	13.54	6.21	42.68
FGSM Probe Attacks	4.65	22.20	4.37	22.04
PGD Probe Attacks	4.37	23.18	4.70	24.81

dings. Further, to illustrate the advantage of the adversarial training in observing the shift in distribution between mated and non-mated comparison scores, we also present the obtained distributions in Fig. 7.10. As it can be noted from Fig. 7.10, the distribution of mated and non-mated comparison becomes very identical to baseline system performance when no attacks are conducted, as shown in Fig. 7.1a.

7.8 Discussion

With the set of all experiments conducted in this work under the open-set protocols for biometric verification using a deep model, we observe that the FRS are generally vulnerable to poisoning/perturbation attacks. Although the deep FRS are sensitive to a different degree based on the degree of poisoning of images, both FGSM and PGD



(a) FRGC with FGSM Attacks (b) FRGC with PGD Attacks (c) FGSM with PGD Attacks

Fig. 7.10 Distribution shift in mated and non-mated comparison scores as a result of the adversarial training in a combined manner using PGD+FGSM samples on cross-poisoning attacks

attacks can adversely affect the false match and false non-match decisions, both of which have a significant operational impact if deployed. We noted that the baseline performance of FRS degrades when the clean data alone (despite capture noises such as bad illumination, pose and expression) is used. The performance of the systems further degrades when the cross-poisoning attacks are carried out, specifically when the attacker can manipulate the images in the enrolment set and probe with images of significant attack degree. Unlike the deep FRS, we also note that the COTS FRS is insensitive to such attacks, but due to limited knowledge on the employed algorithm in COTS, one cannot conclude on what contributes to its robustness.

However, we also note that by simply retraining the entire network with adversarial examples, we can improve the baseline performance of the deep FRS and also make it robust to cross-poisoning attacks. One key benefit of such an approach is the limited overhead on the network design where one can simply reuse the network. While on the other hand, the deep FRS may still remain sensitive to the newer attacks if such examples are not seen by the network during the training phase. Alternatively, one can simply add another layer to the FRS network which can detect adversarial attacks, which is a common practice in presentation attack detection. On the downside of such design is the additional overhead of design of the network and no guarantee that these adversarial sample detection module would scale to newer and unknown attacks. In another direction, stricter constraints can be imposed to eliminate the non-conforming images according to quality standards as defined by ISO/IEC standards—29794-5 [14, 15] should such systems be deployed. Such observations and arguments lead us to critically analyze the deep FRS for various factors and study the generalizing ability to diverse adversarial attacks on FRS. This can be an interesting direction for future works for mitigating the adversarial threats on deep FRS.

7.9 Conclusions and Future Directions

Despite the impressive accuracy obtained with deep models for various face recognition tasks, they are vulnerable to various kinds of attacks. In this chapter, we have presented various adversarial attacks that can negatively impact the biometric performance of face recognition systems. Further, we have chosen two relevant adversarial attacks based on the poisoning of the images at both probe level and enrolment level. The chosen attacks were thoroughly evaluated using a state-of-art face dataset to illustrate the impact of the poisoning attacks on deep network-based face recognition. This chapter specifically illustrated the impact on biometric performance in terms of false match and false non-match decisions when such poisoned data is used for attacks. Further, this chapter also illustrated the use of adversarial examples to make the deep models robust towards such poisoning attacks.

Future works in this direction can also combine the poisoning attacks with the parameter level attacks to verify the impact on biometric performance. Another potential direction is to study the model and parameter protection mechanisms to avoid white-box attacks.

References

1. Agarwal A, Singh R, Vatsa M, Ratha N (2018) 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). IEEE, pp 1–7
2. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, pp 67–74
3. Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: Proceedings—IEEE symposium on security and privacy, pp 39–57
4. Chen W, Zhang Z, Hu X, Wu B (2020) Boosting decision-based black-box adversarial attacks with random sign flip. In: European conference on computer vision. Springer, pp 276–293
5. Dabouei A, Soleymani S, Dawson J, Nasrabadi N (2019) Fast geometrically-perturbed adversarial faces. In: 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1979–1988
6. Deng J, Guo J, Xue N, Zafeiriou S (2019) Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4690–4699
7. Dong Y, Su H, Wu B, Li Z, Liu W, Zhang T, Zhu J (2019) Efficient decision-based black-box adversarial attacks on face recognition, pp 7714–7722
8. Fan Y, Wu B, Li T, Zhang Y, Li M, Li Z, Yang Y (2020) Sparse adversarial attack via perturbation factorization. Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 12367 LNCS, pp 35–50
9. Goel A, Singh A, Agarwal A, Vatsa M, Singh R (2018) Smartbox: benchmarking adversarial detection and mitigation algorithms for face recognition. In: 2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS). IEEE, pp 1–7. IEEE
10. Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv:1412.6572

11. Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: 3rd international conference on learning representations, ICLR 2015-conference track proceedings, pp 1–11
12. Goswami G, Agarwal A, Ratha N, Singh R, Vatsa M (2019) Detecting and mitigating adversarial perturbations for robust face recognition. Int J Comput Vis 127(6):719–742
13. Huang GB, Mattar M, Berg T, Learned-Miller E (2008) Labeled faces in the wild: a database forstudying face recognition in unconstrained environments. In: Workshop on faces in 'Real-Life' images: detection, alignment, and recognition
14. International Organization for Standardization. ISO/IEC TR 29794-5:2010 Information technology–Biometric sample quality–Part 5: Face image data, 2020
15. International Organization for Standardization. ISO/IEC 29794-5:2020 Information technology–Biometric sample quality–Part 5: Face image data, 2020
16. Korshunov P, Marcel S (2018) Deepfakes: a new threat to face recognition? assessment and detection. arXiv:1812.08685
17. Kurakin A, Goodfellow I, Bengio S et al (2016) Adversarial examples in the physical world
18. Massoli FV, Carrara F, Amato G, Falchi F (2021) Detection of face recognition adversarial attacks. Comput Vis Image Underst 202:103103
19. Moosavi-Dezfooli S-M, Fawzi A, Fawzi O, Frossard P (2017) Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1765–1773
20. Moschoglou S, Papaioannou A, Sagonas C, Deng J, Kotsia I, Zafeiriou S (2017) Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 51–59
21. Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A (2017) Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security, pp 506–519
22. Peña A, Serna I, Morales A, Fierrez J, Lapedriza A (2020) Facial expressions as a vulnerability in face recognition. arXiv:2011.08809
23. Phillips PJ, Flynn PJ, Scruggs T, Bowyer KW, Chang J, Hoffman K, Marques J, Min J, Worek W (2005) Overview of the face recognition grand challenge. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 1. IEEE, pp 947–954
24. Qin L, Peng F, Venkatesh S, Ramachandra R, Long M, Busch C (2020) Low visual distortion and robust morphing attacks based on partial face image manipulation. IEEE Trans Biomet Behav Identity Sci
25. Raja K, Ferrara M, Franco A, Spreeuwers L, Batskos I, de Wit Marta Gomez-Barrero F, Scherhag U, Fischer D, Venkatesh S, Singh JM et al (2020) Morphing attack detection–database, evaluation platform and benchmarking. In: IEEE - TIFS
26. Ramachandra R, Busch Christoph (2017) Presentation attack detection methods for face recognition systems: a comprehensive survey. ACM Comput Surveys (CSUR) 50(1):1–37
27. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2016) Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1–11
28. Rozsa A, Rudd EM, Boult TE (2016) Adversarial diversity and hard positive generation. In: IEEE computer society conference on computer vision and pattern recognition workshops, pp 410–417
29. Saha S, Sim T (2020) Is face recognition safe from realizable attacks? In: 2020 IEEE international joint conference on biometrics (IJCB), pp 1–8
30. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
31. Sengupta S, Chen J-C, Castillo C, Patel VM, Chellappa R, Jacobs DW (2016) Frontal to profile face verification in the wild. In: 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1–9

32. Sharif M, Bhagavatula S, Bauer L, Reiter MK (2016) 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, pp 1528–1540
33. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. arXiv:1312.6199
34. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inf Fusion 64:131–148
35. Vakhshiteh F, Nickabadi A, Ramachandra R (2020) Adversarial attacks against face recognition: a comprehensive study. arXiv:2007.11709
36. Venkatesh S, Ramachandra R, Raja K, Busch C (2021) Face morphing attack generation & detection: a comprehensive survey. In: IEEE-TTS
37. Vivek BS, Mopuri KR, Venkatesh Babu R (2018) Gray-box adversarial training. In: Proceedings of the European conference on computer vision (ECCV), pp 203–218
38. Wu D, Xia S-T, Wang Y (2020) Adversarial weight perturbation helps robust generalization. Adv Neural Inf Process Syst 33
39. Xu X, Chen J, Xiao J, Wang Z, Yang Y, Shen HT (2020) Learning optimization-based adversarial perturbations for attacking sequential recognition models. In: Proceedings of the 28th ACM international conference on multimedia, pp 2802–2822
40. Zhao R (2020) Vulnerability of the neural networks against adversarial examples: a survey. arXiv:2011.05976
41. Zhong Y, Deng Weihong (2020) Towards transferable adversarial attack against deep face recognition. IEEE Trans Inf Forensics Secur 16:1452–1466
42. Zhu Z-A, Lu Y-Z, Chiang C-K (2019) Generating adversarial examples by makeup attacks on face recognition. In: 2019 IEEE international conference on image processing (ICIP). IEEE, pp 2516–2520