A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

Toward Practical Adversarial Attacks on Face Verification Systems

Kazuya Kakizaki[1], Taiki Miyagawa[2], Inderjeet Singh[3], Jun Sakuma[4]

Abstract: DNN-based face verification systems are vulnerable to adversarial examples. The previous paper's evaluation protocol (scenario), which we called the probe-dependent attack scenario, was unrealistic. We define a more practical attack scenario, the probe-agnostic attack. We empirically show that these attacks are more challenging than probe-dependent ones. We propose a simple and effective method, PAMTAM, to improve the attack success rate for probe-agnostic attacks. We show that PAMTAM successfully improves the attack success rate in a wide variety of experimental settings.

Keywords: Adversarial example, Face verification, Security.

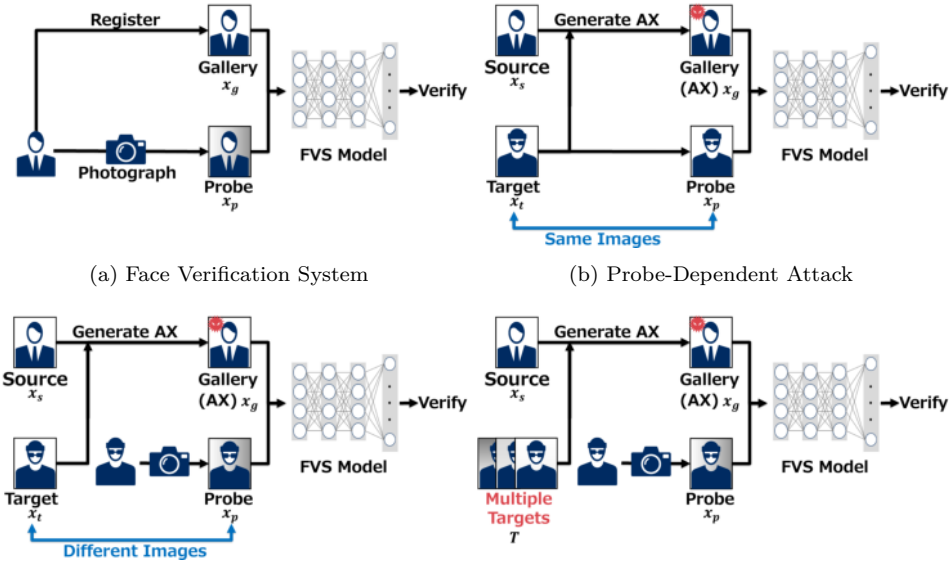
1 Introduction

Face verification systems (FVSs) verify the identity of a person by comparing two face images: *gallery* and *probe* images. The gallery image x_g is registered in the FVS in advance, and the probe image x_p is captured by a camera installed in the FVS at verification time, as shown in Fig. 1a. Recent progress on deep neural networks (DNNs) has significantly improved the performance of FVSs; however, DNNs have been shown to be vulnerable to small, human-imperceptible perturbations to the input data, or *adversarial examples* (AXs) [Sz14], which jeopardize the safety and security of DNN-based FVSs.

There are several studies on adversarial attacks against FVSs [RGB17, ZD20, Do19b]. These studies assume an adversary who generates an AX from images of the victim's and adversary's face (*source image* x_s and *target image* x_t, respectively); the generated AX looks like the victim but is expected to be misidentified as the adversary. Then, they assume an attack scenario in which the adversary can input a generated AX and target image x_t into the DNN in FVSs as a gallery image x_g and probe image x_p, respectively, as shown in Fig. 1b. However, this attack scenario is impractical in real-world settings because the probe images are captured by a camera at verification time [5]. We call this impractical attack scenario ($x_t = x_p$) the *probe-dependent attack*.

[1] NEC Corporation, University of Tsukuba, kazuya1210@nec.com
[2] NEC Corporation, miyagawataik@nec.com
[3] NEC Corporation, inderjeet78@nec.com
[4] University of Tsukuba, RIKEN AIP, jun@cs.tsukuba.ac.jp

[5] If the camera in the FVS is under the control of the adversary, then the adversary can input a generated AX directly into the FVS as the probe image. In our paper, however, we assume



(a) Face Verification System

(b) Probe-Dependent Attack

(c) Probe-Agnostic Attack

(d) Proposed Method (PAMTAM)

Fig. 1: (a) Face verification systems (FVSs) verify the identity of a person by comparing two face images: *gallery* image x_g, which is registered in the FVS in advance, and *probe* image x_p, which is captured by the camera installed in the FVS at the verification time. (b) Probe-Dependent Attack assumes that the adversary can input a generated AX and target image into the DNN in FVSs as a gallery image and probe image ($x_t = x_p$). (c) Probe-Agnostic Attack assumes that the adversary cannot input a target image as a probe image ($x_t \neq x_p$). (d) Our method, PAMTAM, generates AXs using multiple target face images T. Note that the aforementioned settings are different from the presentation attack [Hu19], which is outside the scope of the present paper.

In this paper, we consider a more practical but challenging attack scenario, the *probe-agnostic* attack, as shown in Fig. 1c. We do not assume that $x_t = x_p$; thus, there generally exists a domain gap between x_t and x_p depending on when, where, and how the images are captured (*capturing conditions*), e.g., the illumination conditions, head poses, and image resolution. To the best of our knowledge, the probe-agnostic attack is yet to be explored in the literature and is important for assessing the true risk of practical adversarial attacks against FVSs.

The difficulty with the probe-agnostic attack comes from the domain gap between x_t and x_p due to the different capturing conditions. To address this problem, we propose a simple but effective method for increasing the attack success rate for probe-agnostic attacks: the Probe-Agnostic Multiple Target Method (PAMTAM). PAMTAM makes *arbitrary* attack methods robust to variable capturing conditions,

that the adversary cannot hack the FVS, which is the case, e.g., the automated face recognition gates at airports.

irrespective of white- or black-box attacks. PAMTAM generates AXs so that it approaches, on average, *multiple* target images in the feature space to make the features robust to domain gaps, as shown in Fig. 1d. We empirically show that PAMTAM successfully increases the attack success rates of 2 widely used attacks on 3 databases with 8 different model combinations, attaining a maximum relative recovery of 83.3%.

Our contribution is twofold. First, we formulate the probe-agnostic attack, which is yet to be explored in the literature and is important for assessing the true risk of practical adversarial attacks against FVSs. Next, we propose PAMTAM, which makes arbitrary attack methods robust to variable capturing conditions. We empirically show that PAMTAM successfully improves the attack success rate under a wide variety of conditions.

1.1 Related Work

All the following studies focus on the AXs against FVSs but follow the probe-dependent scenario. In the white-box setting, the attacker can access the network structure and parameters of the target DNN. [Sa16] is the first to show that DNN-based feature extractors, not only classifiers, are vulnerable to AXs. [RGB17] proposed LOTS that generates AX that is close to the target face image in the feature space. [ZD19] proposed Iterative Feature Target Gradient Sign Method (IFTGSM), which iteratively updates AX with a gradient sign of the gradient. [SWY18, DZJ19] leveraged Generative Adversarial Networks (GANs) to generate AXs with high perceptual quality. In contrast, the black-box attack assumes that the attacker cannot access the network structure or parameters of the target DNN. [Do19b] proposed a query-based attack method, where the attacker could send queries to FVSs and see the outputs. The query-based attack can relatively high attack success rates but can be easily detected because a number of queries are necessary to generate AXs, causing a suspiciously large amount of accesses to the target FVS. [ZD20] used surrogate models to generate AXs without queries. The authors proposed the dropout face attacking network (DFANet) to enhance transferability. They also showed that [Li17, Xi19, Do18], originally used for classifiers, are effective even for feature extractors.

2 Preliminaries

Face verification systems. Face verification is a task to determine whether two face images are derived from the same identity. Modern FVSs use DNN-based feature extractors [De19, Wa18]. Let $\mathscr{X}$ be a set of images with height $H \in \mathbb{N}$, width $W \in \mathbb{N}$, and the number of channels $C \in \mathbb{N}$, i.e., $\mathscr{X} = \{0, 1, ..., 255\}^{H \times W \times C}$. Let $f : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^d$ be a feature extractor, where $d \in \mathbb{N}$ is the feature dimension. We define a function **Ver**, which represents the internal processes of FVSs, as a

mapping from two images ($x_1, x_2 \in \mathscr{X}$) to a binary set ($\{\text{Verified}, \text{NotVerified}\}$):

$$\mathbf{Ver}_{f,\alpha}(x_1, x_2) = \begin{cases} \text{Verified} & (\text{if } \mathbf{dist}(f(x_1),\, f(x_2)) \leq \alpha) \\ \text{Not Verified} & (\text{otherwise}), \end{cases} \tag{1}$$

where $\alpha \in \mathbb{R}_{\geq 0}$ is a threshold, and $\mathbf{dist} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is an arbitrary distance function of feature vectors. Typically, the cosine similarity or L^2 norm are used for face verification. We use the latter in the present paper, but the extensions to other distance functions are straightforward.

Our primary focus is the FVSs which involve the following two steps (See also Fig. 1a):

1. **Registration.** A user registers her or his face image (gallery image $x_g \in \mathscr{X}$) with the FVS. The FVS stores $f(x_g)$ in the gallery set.
2. **Verification.** At the verification phase, the FVS takes a photo of the user (probe image $x_p \in \mathscr{X}$) with the internal camera, which is sometimes invisible to the user. Then, the FVS computes $f(x_p)$ and runs Eq. (1) to investigate whether the two identities are the same.

Adversarial attacks against face verification systems. We assume that the gallery image x_g is an AX and the probe image x_p is not, as in Introduction, although there are two other possibilities in principle: (i) x_p is an AX, and (ii) both x_g and x_p are AXs. We can see that both (i) and (ii) are possible but infeasible, because the attacker is required to hack the FVS to input an AX directly to it. A potential solution is the physical adversarial attack [Sh16], but this is out of the scope of the present paper. Therefore, the attacker's goal is formally summarized into the problem of finding adversarial noise $\delta \in \mathbb{R}^{H \times W \times C}$ such that

$$\mathbf{dist}(f(x_g = x_s + \delta), f(x_p)) \leq \alpha \tag{2}$$

$$\|\delta\|_\infty \leq \varepsilon, \tag{3}$$

where $||\cdot||_\infty$ denotes the L^∞ (max) norm. Eq. (3) restricts the size of δ and ensures that the noise is imperceptible to humans. In general, δ is a function of x_s and x_t.

3 PAMTAM

Probe-dependent and probe-agnostic attacks. A common way to generate δ is to define an objective function and minimize it. In the probe-dependent attack, $x_t = x_p$ and $\delta = \delta(x_s, x_p)$ (Fig. 1b); therefore, Eq. (2) can be achieved by minimizing the objective function

$$J(x_s + \delta, x_p, f) = \|f(x_s + \delta) - f(x_p)\|_2^2 \tag{4}$$

with respect to δ. The adversarial noise thus obtained, denoted by $\delta^*(x_s, x_p)$, deceives the target FVS more easily than in the probe-agnostic scenario because $\delta^*(x_s, x_p)$ has the prior knowledge of x_p. In comparison, the probe-agnostic attack assumes $x_t \neq x_p$ and $\delta = \delta(x_s, x_t)$ (Fig. 1c); therefore, the objective function is

$$J(x_s + \delta, x_t, f) = \|f(x_s + \delta) - f(x_t)\|_2^2. \tag{5}$$

The solution $\delta^*(x_s, x_t)$ has no prior knowledge of x_p and is likely to overfit to x_t; therefore, the AX $x_g = x_s + \delta^*(x_s, x_t)$ has no guarantee of being misidentified as x_p if the domain gap between the two (x_t and x_p) is large. In fact, we empirically show in Section 4 the degradation from $\delta^*(x_s, x_p)$ to $\delta^*(x_s, x_t)$; probe-agnostic attacks are more challenging than probe-dependent attacks.

Proposed method. To achieve better attack success rates in the probe-agnostic scenario, we propose *diversifying the target image* x_t, introducing the *target image set* $T = \{x_t^i \in \mathscr{X} | i = 1, ..., |T|\}$, and modifying the objective function (5) as

$$J(x_s + \delta, T, f) = \frac{1}{|T|} \sum_{x_t \in T} \|f(x_s + \delta) - f(x_t)\|_2^2. \tag{6}$$

The target image set consists of facial images of the attacker, which should cover the domain gaps between x_t and x_p, such as different head poses, illumination conditions, image resolutions, facial expressions, and makeup. In fact, our experiment shows that a larger T enhances the attack success rate (Section 4). Note that it is easy to increase the sample size of T in practice, compared with source and probe images, because the attacker can take selfies under arbitrary conditions. We set $|T| = 5$ in our experiments unless otherwise noted.

A motivation of Eq. (6) comes from the recent studies showing that the diversity of the input images improves transferability [Do19a, Xi19]; however, no previous work has explored it to attack FVSs especially in the probe-agnostic scenario. Moreover, a crucial difference between our method and [Do19a, Xi19] is that the latter uses automatic, mechanical transformations for the input diversity (random resizing, random padding, and translation). However, such transformations are not sufficient to fill the large, complex domain gaps. In addition, their transormations are applied to the *source image*, while our method diversify the *target image*, to adapt the adversarial example to the probe-agnostic scenario.

The proposed method, *PAMTAM*, is widely applicable to arbitrary objective functions for (2) and arbitrary optimization methods, e.g., [ZD19, RGB17, Sa16, Do19b] . PAMTAM does not even depend on whether the attack is white-box or black-box.

4 Experiment

In this section, we demonstrate that probe-agnostic attacks are more challenging than probe-dependent attacks, as mentioned in Section 3. We then show that PAM-

TAM successfully improves the attack success rate. Our experiments are based on 2 attack methods on 3 datasets under 8 different conditions, as explained below. We focused on three domain gaps, head poses, illumination conditions, and image resolutions, which are likely to occur when we use an actual FVS.

To simulate the most realistic situations, all the experiments assumesd that the attacker cannot access the FVS model. First, we trained a DNN model (*FVS model*) on a training dataset (*FVS dataset*). Second, we trained another DNN model (*surrogate model*) on another dataset (*surrogate dataset*) to perform the surrogate model attack. Third, we sampled (i) source-probe doublets (x_s, x_p) and (ii) source-probe-target triplets (x_s, x_p, x_t) from yet another dataset (*material dataset*), which should have no intersection with the FVS or surrogate datasets. Note that this step distinguishes our experiments from those in preceding papers. (i) and (ii) were used for probe-dependent and probe-agnostic attacks, respectively. x_t in (ii) was replaced with T when PAMTAM was used. Fourth, using (i) and (ii), we generated AXs that deceive the surrogate model. Finally, using the AXs thus generated, we evaluated their attack success rates on the FVS model. The evaluation measure was the attack success rate, i.e., the proportion of the AXs matched to the probe images.

$$\sum_{x_s, x_p, \delta \in D} \frac{\mathbb{1}(\mathbf{Ver}_{f,\alpha}(x_s + \delta, x_p) = \text{Verified})}{|D|}, \tag{7}$$

where the test set D was defined as $\{(x_s^i, x_p^i, \delta(x_s^i, x_p^i))\}_{i=1}^{|D|}$, $\{(x_s^i, x_p^i, x_t^i, \delta(x_s^i, x_t^i))\}_{i=1}^{|D|}$, or $\{(x_s^i, x_p^i, T^i, \delta(x_s^i, T^i))\}_{i=1}^{|D|}$ for the probe-dependent attacks, probe-agnostic attacks, and PAMTAM, respectively ($|D| = 200$ for our experiments). The verification threshold α of the FVS model was determined to achieve the best verification accuracy on the LFW dataset [Hu07]. All the FVS models in our experiments achieved a verification accuracy of 98% or higher.

Surrogate and FVS datasets and models. Though not essential, we slightly modified the objective functions (4), (5), and (6) to improve the base attack success rate of all the methods. Following [Li17, Xi19, ZD20], we introduced multiple surrogate models ($F = \{f_i\}_{i=1}^{|F|}$) and stochastic transformations τ of $x_s + \delta$ and took the average over F and τ. F is defined in Tab. 1 ($|F| = 5$). τ includes random resizing and padding.

Our experiments used 8 combinations of the surrogate dataset, surrogate models, FVS dataset, and FVS model (Tab. 1). We used seven network architectures: residual network (R50, R100) [He16], inception residual network (IR50, IR100) [Sz17], squeeze-and-excitation inception residual network (SE50, SE100), and MobileFaceNet (MOB) [Ch18], each of which was attached with a state-of-the-art loss function (ArcFace (Arc) [De19] or CosFace (Cos) [Wa18]). We used two datasets: MS1MV2 (MS) [De19] and VGGFace2 (VGG) [Ca18].

Tab. 1: **Surrogate and FVS datasets and models.** We use four conditions (I, II, III, and IV).

	Surrogate			FVS		
	Data	Loss	Architecture	Data	Loss	Architecture
I	MS	Arc	R100,R50,IR100, IR50,SE50	VGG	Cos	MOB
II	MS	Arc	R100,R50,IR100,IR50,SE50	VGG	Cos	SE100
III	MS	Cos	R100,R50,IR100,IR50,SE50	VGG	Arc	MOB
IV	MS	Cos	R100,R50,IR100,IR50,SE50	VGG	Arc	SE100
V	VGG	Arc	R100,R50,IR100, IR50,SE50	MS	Cos	MOB
VI	VGG	Arc	R100,R50,IR100,IR50,SE50	MS	Cos	SE100
VII	VGG	Cos	R100,R50,IR100,IR50,SE50	MS	Arc	MOB
VIII	VGG	Cos	R100,R50,IR100,IR50,SE50	MS	Arc	SE100

Attack methods. We used two standard attack methods: the Sabour's attack (SAB) [Sa16] and the iterative feature target gradient sign method (IFTGSM) [ZD19]. SAB minimized Eq. (2) under the constraint Eq. (3) by using a box-constrained L-BFGS. IFTGSM iteratively updated δ as

$$\delta^{i+1} = C_\varepsilon(\delta_i - \mathrm{sign}(\nabla_{x_s+\delta^i} J(x_s, x_p, \delta, f))), \tag{8}$$

where $C_\varepsilon(\cdot)$ is a clipping function with a max radius ε, and $\mathrm{sign}(\cdot)$ is a sign function. In our experiments, the maximum perturbation ε was 10 in terms of the L^∞ norm; therefore, the perturbation to the pixel range was at most $10/255 \simeq 3.9\%$.

Material datasets. We use Head Pose Image Database [GHC04], Extended Yale Face Database B [LHK05], and VGGFace2 [Ca18]. They allow us to simulate various types of the domain gaps: head poses (Head Pose Image Database); illumination conditions (Extended Yale Face Database B); and combinations of head poses, illumination conditions, and image resolutions (VGGFace2). Head Pose Image Database consists of 2790 color facial images of 15 individuals with variations of vertical and horizontal face angles. These angles are expressed from -90 degrees to 90 degrees. Extended Yale Face Database B contains 16128 grayscale facial images of 28 individuals under 9 poses and 64 illumination conditions. We only use frontal facial images to fix the head pose. VGGFace2 consists of 3.31 million color facial images of 9131 persons, which covers a wide variety of head poses, illumination conditions, and image resolutions. Note that we also use VGGFace2 to train the FVS datasets, but there is no duplication with the material dataset.

4.1 Results

All the results are summarized in Fig. 2. PD and PA are short for probe-dependent and probe-agnostic. PD is the baseline and attained comparable performances with

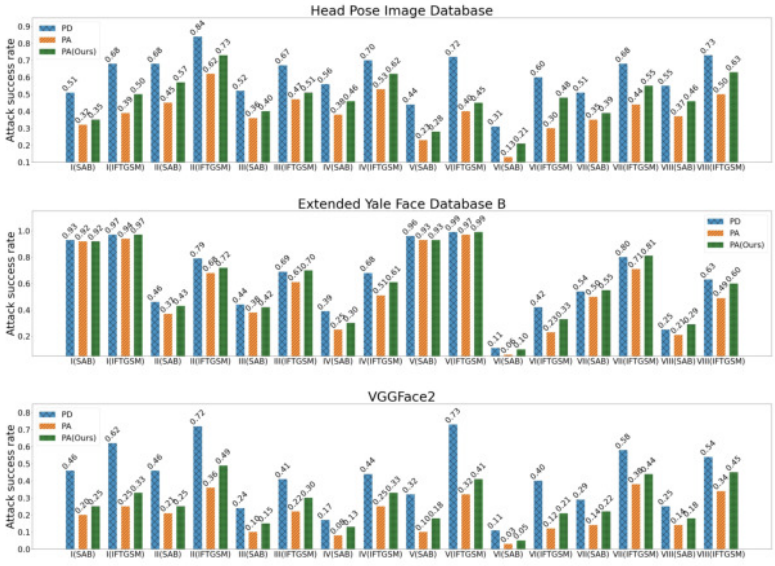


Fig. 2: **Degradation (PD→PA) and recovery (PA→PA(Ours)) of attack success rate.** Three subfigures correspond to three material datasets. PD and PA are short for probe-dependent and probe-agnostic. PA(Ours) is PAMTAM. I, II, III, IV, V, VI, VII, and VIII correspond to those in Tab. 1. Note that the numbers below the third decimal place are omitted.

a recent paper [ZD20][6]. PD achieved the best performance compared with the others PA and PA(Ours), and we found a significant degradation from PD to PA; *PA was more challenging than PD.* The attack success rates decreased by up to 56.4% for the Head Pose Image Database, 45.4% for Extended Yale Face Database B, and 73.9% for VGGFace2. The degradation of VGGFace2 was relatively large because the domain gap between x_t and x_p was larger than the other two. PAMTAM successfully increased the rates under almost all the conditions. The rates increased by up to 61.6% for the Head Pose Image Database, 75.0% for Extended Yale Face Database B, and 83.3% for VGGFace2. Fig. 3 shows PAMTAM's dependence on $|\boldsymbol{T}|$ (evaluated on the VGGFace2 material dataset). We confirmed that large sample sizes help to enhance the attack success rate. The performance gain gradually saturated.

5 Conclusion

This paper considered adversarial attack against FVSs. We defined a more practical attack scenario (probe-agnostic attacks) than that in the previous paper (probe-dependent attacks). We empirically showed that probe-agnostic attacks are more

[6] Note that in face verification, attack success rates fluctuate significantly, depending on the DNN model and dataset (e.g., see [ZD20]).

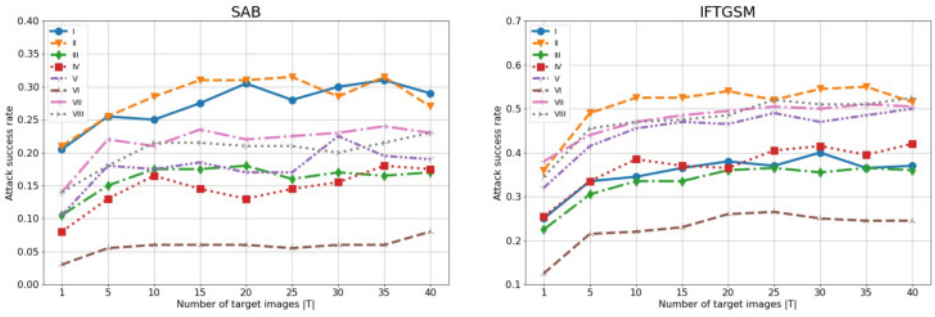


Fig. 3: **Dependence of PAMTAM on $|T|$.** Two subfigures correspond to the two attacks. I, II, III, IV, V, VI, VII, and VIII correspond to those in Tab. 1.

challenging than probe-dependent ones. The results above suggest that previous papers have overestimated the risk of AXs, especially when the domain gaps between x_p and x_t are large. We proposed PAMTAM, which successfully increase the attack success rate of probe-agnostic attacks. We conclude that we should evaluate not only probe-dependent attacks but also probe-agnostic ones under practical domain gaps to correctly capture the threat of AXs to FVSs.

Acknowledgment

We would like to thank to Toshinori Araki and anonymous reviewers for insightful discussions.

References

[Ca18] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; Zisserman, A.: VGGFace2: A dataset for recognising faces across pose and age. In: International Conference on Automatic Face and Gesture Recognition. 2018.

[Ch18] Chen, Sheng; Liu, Yang; Gao, Xiang; Han, Zhen: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition. Springer, pp. 428–438, 2018.

[De19] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699, 2019.

[Do18] Dong, Yinpeng; Liao, Fangzhou; Pang, Tianyu; Su, Hang; Zhu, Jun; Hu, Xiaolin; Li, Jianguo: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193, 2018.

[Do19a] Dong, Yinpeng; Pang, Tianyu; Su, Hang; Zhu, Jun: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4312–4321, 2019.

[Do19b] Dong, Yinpeng; Su, Hang; Wu, Baoyuan; Li, Zhifeng; Liu, Wei; Zhang, Tong; Zhu, Jun: Efficient decision-based black-box adversarial attacks on face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7714–7722, 2019.

[DZJ19] Deb, Debayan; Zhang, Jianbang; Jain, Anil K: Advfaces: Adversarial face synthesis. arXiv preprint arXiv:1908.05008, 2019.

[GHC04] Gourier, Nicolas; Hall, Daniela; Crowley, James L: Estimating face orientation from robust detection of salient facial features. In: ICPR International Workshop on Visual Observation of Deictic Gestures. Citeseer, 2004.

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778, 2016.

[Hu07] Huang, Gary B.; Ramesh, Manu; Berg, Tamara; Learned-Miller, Erik: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[Hu19] Husseis, Anas; Liu-Jimenez, Judith; Goicoechea-Telleria, Ines; Sanchez-Reillo, Raul: A Survey in Presentation Attack and Presentation Attack Detection. In: 2019 International Carnahan Conference on Security Technology (ICCST). pp. 1–13, 2019.

[LHK05] Lee, Kuang-Chih; Ho, Jeffrey; Kriegman, David J: Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on pattern analysis and machine intelligence, 27(5):684–698, 2005.

[Li17] Liu, Yanpei; Chen, Xinyun; Liu, Chang; Song, Dawn: Delving into transferable adversarial examples and black-box attacks. International Conference on Learning Representations, 2017.

[RGB17] Rozsa, Andras; Günther, Manuel; Boult, Terranee E: LOTS about attacking deep features. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 168–176, 2017.

[Sa16] Sabour, Sara; Cao, Yanshuai; Faghri, Fartash; Fleet, David J: Adversarial manipulation of deep representations. International Conference on Learning Representations, 2016.

[Sh16] Sharif, Mahmood; Bhagavatula, Sruti; Bauer, Lujo; Reiter, Michael K: 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, 2016.

[SWY18] Song, Qing; Wu, Yingqi; Yang, Lu: Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. arXiv preprint arXiv:1811.12026, 2018.

[Sz14] Szegedy, Christian; Zaremba, Wojciech; Sutskever, Ilya; Bruna, Joan; Erhan, Dumitru; Goodfellow, Ian; Fergus, Rob: Intriguing properties of neural networks. International Conference on Learning Representations, 2014.

[Sz17] Szegedy, Christian; Ioffe, Sergey; Vanhoucke, Vincent; Alemi, Alexander: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. volume 31, 2017.

[Wa18] Wang, Hao; Wang, Yitong; Zhou, Zheng; Ji, Xing; Gong, Dihong; Zhou, Jingchao; Li, Zhifeng; Liu, Wei: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274, 2018.

[Xi19] Xie, Cihang; Zhang, Zhishuai; Zhou, Yuyin; Bai, Song; Wang, Jianyu; Ren, Zhou; Yuille, Alan L: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2730–2739, 2019.

[ZD19] Zhong, Yaoyao; Deng, Weihong: Adversarial learning with margin-based triplet embedding regularization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6549–6558, 2019.

[ZD20] Zhong, Yaoyao; Deng, Weihong: Towards Transferable Adversarial Attack against Deep Face Recognition. arXiv preprint arXiv:2004.05790, 2020.