Multimedia Tools and Applications (2023) 82:7443–7461
https://doi.org/10.1007/s11042-022-13641-1

ApaNet: adversarial perturbations alleviation network for face verification

Guangling Sun[1] · Haoqi Hu[1] · Yuying Su[1] · Qi Liu[1] · Xiaofeng Lu[1]

Received: 1 March 2021 / Revised: 21 March 2022 / Accepted: 2 August 2022 /
Published online: 23 August 2022

Abstract

Albeit Deep neural networks (DNNs) are widely used in computer vision, natural language processing and speech recognition, they have been discovered to be fragile to adversarial attacks. Specifically, in computer vision, an attacker can easily deceive DNNs by contaminating an input image with perturbations imperceptible to humans. As one of the important vision tasks, face verification is also subject to adversarial attack. Thus, in this paper, we focus on defending against the adversarial attack for face verification to mitigate the potential risk. We learn a network via an implementation of stacked residual blocks, namely adversarial perturbations alleviation network (ApaNet), to alleviate latent adversarial perturbations hidden in the input facial image. During the supervised learning of ApaNet, only the Labeled Faces in the Wild (LFW) is used as the training set, and the legitimate examples and corresponding adversarial examples produced by projected gradient descent algorithm compose supervision and inputs respectively. By leveraging the middle and high layer's activation of FaceNet, the discrepancy between an image output by ApaNet and the supervision is calculated as the loss function to optimize ApaNet. Empirical experiment results on the LFW, YouTube Faces DB and CASIA-FaceV5 confirm the effectiveness of the proposed defender against some representative white-box and black-box adversarial attacks. Also, experimental results show the superiority performance of the ApaNet as comparing with several currently available techniques.

Keywords Deep neural network · Face verification · Adversarial example · Adversarial perturbations alleviation network

✉ Xiaofeng Lu
luxiaofeng@shu.edu.cn

[1] Shanghai University, School of Communication and Information Engineering, 99 Shangda Road, Baoshan District, Shanghai 200444, China

1 Introduction

Recently, deep neural networks (DNNs) have achieved remarkable performance in computer vision tasks, including sign language recognition [35], salient object detection [17], anomaly crowded detection [18] and so on. As one of the important computer vision tasks, facial biometric research based on DNNs has also greatly advanced [41] and related applications have been deployed in surveillance and access control, such as payment, public access, criminal verification [1].

However, Szegedy et al. [39] first discover that elaborately designed adversarial examples, which are imperceptible to humans, can easily deceive DNNs. Since then, numerous of attacking methods have been proposed in literature [30, 48]. Similarly, facial analysis applications based on DNNs also tend to be brittle to adversarial examples. For instance, Rozsa et al. [31] propose fast flipping attribute attacking to alter the result of the facial attribute recognition. Mirjalili and Ross [24] perturb a face image such that sole gender attribute is flipped whereas other biometric information remains unchanged. Chhabra et al. [3] also design adversarial perturbations to alter selected attributes while preserving identity information and visual content.

With the advent of the Era of Internet and cloud technology, the digitization of users' personal information has become an irresistible trend [20]. The resulting user privacy issues have been widely concerned. The European Community has issued a new regulation, named General Data Protection Regulation (GDPR), to ensure users have greater control over the data they provide. Facial image is one of the most important personal information for users and is usually used for identity verification in face recognition system [27]. The problem of face image leakage will aggravate the threat of adversarial attack to the face recognition models.

To counteract the attacks, a plethora of defending approaches have emerged accordingly which roughly fall into four categories: The first is adversarial training [22], as a type of data augmentation scheme, to boost the model's robustness to adversarial perturbations. The second is defensive distillation. Papernot et al. [28] train the classifier in a certain way such that it is nearly impossible for gradient based attacks to generate adversarial examples directly on the network. The third is adversarial examples detection. Fan et al. [7] present an integrated detection framework involving statistical detector and Gaussian noise injection detector. Massoli et al. [23] propose a facial adversarial detection in which the attacked model typically only acts as the feature extractor. The fourth is perturbation cleaning before the analysis by the model. Xie et al. [45] use randomization as a defender. The input images are randomly resized and added random padding prior to the target network to reduce the influence of adversarial perturbations. Jia et al. [16] design an image compression model composed of a compression module and a reconstruction module to purify the adversarial perturbations.

The adversarial examples of the two attacks are shown in Fig. 1. Then, we learn an adversarial perturbations alleviation network (ApaNet) via an implementation of stacked residual networks, to mitigate adversarial perturbations injected into the input image. The third row in Fig. 1 demonstrates the results obtained by our proposed ApaNet. Specifically, given pairs of legitimate images and its adversarial version produced by PGD, as supervision and input respectively, the ApaNet is supervised learned by minimizing a loss function, in which representations of FaceNet, are leveraged to measure the distance between the image output by ApaNet and the supervision legitimate image (see Fig. 2). The motivation behind our proposed loss function is that middle and high layer's feature maps are more related to the ultimate task performance and convey more semantic features. Both training and testing of the

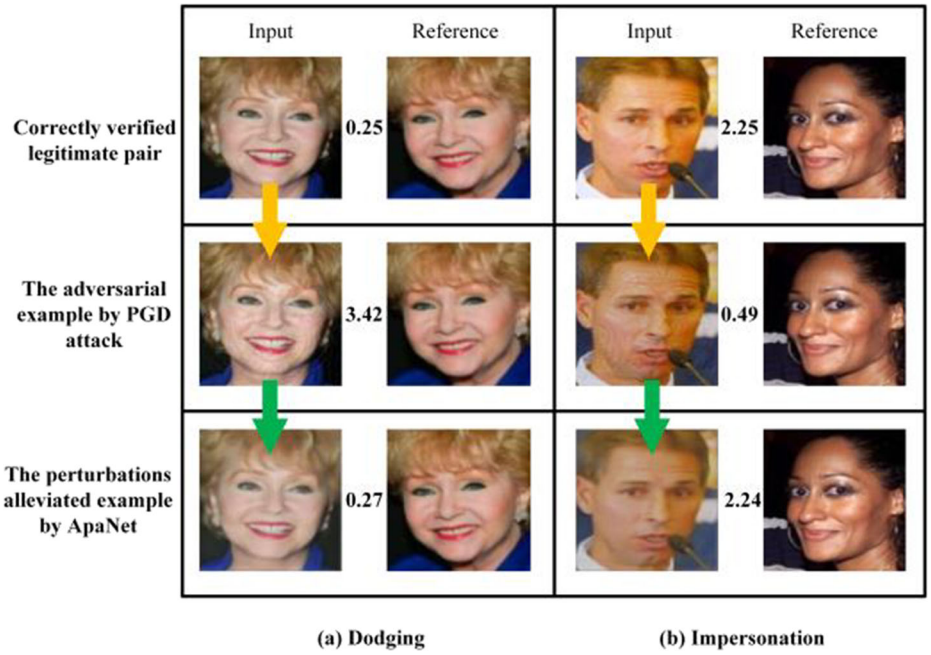

Fig. 1 Adversarial examples and perturbations alleviated examples of dodging attack (**a**) and impersonation attack (**b**). The value between two images measures their similarities. A smaller value tends to give one identity decision whereas a larger value tends to give two distinct identities decision. The validation threshold = 1.1

ApaNet are efficient and the empirical results confirm that the network is capable of counteracting both white-box and black-box attacks.

We summarize our contributions as follows:

1) We design ApaNet, which is a network with an implementation of stacked residual blocks to alleviate the adversarial perturbations. Supervised learning of ApaNet is efficient and stable with a moderate size of training examples.
2) We propose a novel loss function to optimize ApaNet. The middle and high representations of FaceNet, the target network, are leveraged to measure discrepancy between the output image of ApaNet and the supervision legitimate image for the loss function.

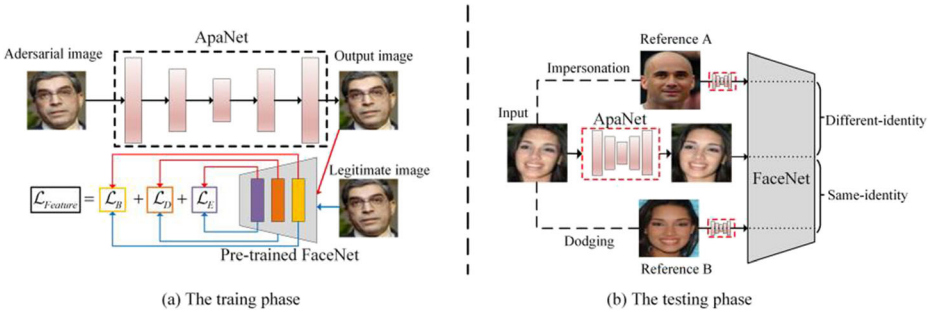

Fig. 2 The overview of proposed defense framework based on ApaNet. (**a**) shows the training phase of ApaNet with the assistance of FaceNet. (**b**) shows ApaNet protects FaceNet against impersonation and dodging attacks during the test phase

3) We conduct comprehensive experiments to verify the effectiveness of ApaNet as defender. The empirical results confirm that it is superior to compared methods for defending both white-box attacks and black-box attacks on the Labeled Faces in the Wild (LFW) [15], YouTube Faces DB [44] and CASIA-FaceV5 [47].

2 Related work

2.1 Deep face recognition

The development of face recognition has been greatly facilitated by DNNs. DeepFace [41] define face recognition as a multi-class classification problem and use a DNNs model trained with softmax loss to identify faces. FaceNet [33] is trained by minimizing a triplet loss and outputs embeddings within a Euclidean space to measure face similarity. CosFace [43] propose a large margin cosine loss to maximize a cosine margin term in the angular space for face recognition task. PocketNet [2] is an extremely lightweight and accurate face recognition system which employed a multi-step knowledge distillation to enhance its verification performance. D-FES [36] use the recurrent neural network to detect human emotions based on the facial lips structure which can accurately track and classify face emotions in a real-time environment. During the Covid-19 pandemic, Face Mask Detection System [26] is designed for verifying whether a person wears a mask which used model pruning to implement embedded deployment.

2.2 Adversarial attacks

Goodfellow et al. [9] propose an efficient and single step attack, Fast Gradient Sign Method (FGSM). Moosavi Dezfooli [25] propose DeepFool which compute the minimal distortion required to force the target model to give a false output. Carlini & Wagner (CW) [37] use an optimization algorithm to tailor adversarial attacks. Madry et al. [22] propose PGD, an iterative attack which is widely used in adversarial training or to evaluate the adversarial robustness of model. Papernot et al. [29] find that adversarial examples are transferable from model to model. Attackers implement black-box transfer-based attacks by training their own substitute model and crafting adversarial examples against the substitute. Without accessing to DNN's parameters, Li et al. [19] estimate a probability density distribution for a neighborhood of the input such that a sample drawn from it is almost adversarial (NATTACK). Sharif et al. [34] develop a physical attack by printing an eyeglass frame to fool the face recognition system in real world. Duan et al. [6] camouflage physical-world adversarial images with a natural style that is invisible to human. Dabouei et al. [4] present a fast landmark manipulation method based on the geometric features of faces to form adversarial examples. Rozsa et al. [32] introduce the layer-wise origin-target synthesis that imitates the deep features of the target to produce adversarial examples (LOTS).

2.3 Adversarial defenses

Defense on neural networks is much more challenging compared with attacks. We summarize some ideas of current approaches to defense and compare them with our work as show in Table 1.

Table 1 A comparison of referred defense methods

Types of defenses	Retraining target model	Defense-aware attack	Generalization
Adversarial training	Required	Defensible	Well
Defensive distillation	Required	Indefensible	Well
Detection	Not required	Defensible	Poor
Perturbation cleaning	Not required	Defensible	Well

Adversarial training One idea of defending against adversarial examples is to train a better classifier. Madry et al. [22] use adversarial example for iterative training to improve the robustness of the model. Tramèr et al. [42] propose "ensemble adversarial training". Additional adversarial examples produced from external pre-trained models are used to enrich training data so as to improve the robustness to the transferred examples. Xie et al. [46] find that the ReLU activation function weakens adversarial learning, and propose smooth adversarial learning which can improve the robustness of the model without reducing the accuracy.

While adversarial training is regarded as one of the most effective defenders, its relatively high complexity remains an incompletely settled problem. Our approach is orthogonal to this branch of work. ApaNet is an additional defense framework that does not require modification to the target classifier.

Defensive distillation Papernot et al. [28] prove that the model's sensitivity to small perturbations can be suppressed by high temperature softmax and proposed defensive distillation mechanism accordingly. The distillation model hides the gradient between the pre-softmax layer (logits) and softmax outputs which defend against gradient-based attacks. However, attackers can still evade defensive distillation by transfer-based attacks or calculating gradients using logits instead of softmax output.

We argue that in defense-aware attack where the attacker knows the parameters of the defense network, it is very difficult to prevent adversaries from crafting adversarial examples. Instead, as a perturbation alleviation network, ApaNet is still defensive against defense-aware attacks (in Section 4.3.3).

Detection Another idea of defense is to detect adversarial examples before data is entered into the model. Deb et al. [5] propose "FaceGuard", a self-supervised adversarial defense framework which can detect adversarial face images without training adversarial examples. Hu et al. [14] propose a two-stream method by analyzing the frame-level and temporality-level information to detect compressed deepfake video. Liao et al. [21] design an order forensics framework for detecting image operator chain which can capture both tampering artifact evidence and local noise residual evidence. Goswami et al. [11] study a methodology for automatic attack detection using the response from hidden layers of the DNNs and a technique of selective dropout in the DNNs to diminish the effect of adversarial attacks. However, the above detectors do not generalize well across different dataset and different attack generation processes.

Perturbation cleaning Perturbation cleaning methods remove any possible adversarial perturbations from the image in the input phase. Guo et al. [13] use image transformations before feeding the adversarial inputs into the system, such as bit-depth reduction, JPEG compression, etc. Goel et al. [8] develope a SmartBox for benchmarking the performance of adversarial attack detection and mitigation algorithm on face recognition task. The above methods can effectively

remove the disturbance by irreversibly deforming the image, and then inevitably reduce the performance of the target model. For face validation model, the input data is generally the face image with high definition and rich texture features. The existing perturbation cleaning methods perform poorly in the face verification task which destroy the important facial information easily.

ApaNet is also a defense method by cleaning perturbation. Contrary to previous work, we do not use the distance in pixels between adversarial images and legitimate images as the supervision information. Instead, we use the deep feature representation of the target model to learn a reasonable mapping from the adversarial images to the legitimate ones, which can maintain the baseline accuracy of the target model. Further, since the perturbation alleviated images fit the legal input distribution of the target model, ApaNet has good generalization in diverse adversarial attacks and datasets.

3 Methodology

3.1 An overview of the proposed ApaNet

The architecture of defense method for face verification includes a generative network called ApaNet and a pre-trained FaceNet, as shown in Fig. 2. ApaNet is a fine image reconstruction network, which aims to alleviate latent adversarial perturbations on face images. FaceNet is an excellent face verification model as the instance of such DNNs without loss of generality. In our work, FaceNet is taken as the target network protected by ApaNet, and its weight parameters did not change in the training or test phase.

In the training phase, the ApaNet which is in the service of mitigating adversarial perturbations, is learned with the aid of FaceNet. As the parameters of ApaNet are optimized via supervised learning, in which the supervision information is legitimate image and corresponding adversarial image, the network intrinsically learns a mapping from adversarial image to legitimate image. It is undoubtedly that designing an effective loss function evaluating the discrepancy between an output image and legitimate image is substantially important for ApaNet. Thus, by leveraging the middle and high layer's activation of FaceNet, we propose a loss function through comparing the distance between the multi-layers' activation for the output image and legitimate image respectively. In our work, the adversarial images used during training are produced by attacking FaceNet using PGD algorithm.

In the testing phase, each face images are cleaned by ApaNet and then input to FaceNet for identity verification. As shown in Fig. 2(b), the input image awaiting verification is forged as the identity of the reference 'A' by impersonation attack. After pre-cleaning by ApaNet, FaceNet can correctly identify the identity of the person in the image. Similarly, under dodging attacks, the image to be verified and reference 'B' is judged by FaceNet as different identities. After ApaNet cleaning, two images with the same identity will be correctly verified. It should be emphasized that ApaNet also performs the same input processing on legitimate images which will not affect its verification accuracy in FaceNet.

3.2 The structure of FaceNet and ApaNet

In this section, the structure of the selected target model FaceNet and the proposed ApaNet are described in detail.

FaceNet FaceNet is a unified system of including a batch input layer, a network followed by a L_2 normalization layer, and outputs an embedding as a facial descriptor. The major contribution of FaceNet is the triplet loss employed to minimize intra-class distance and maximum inter-class distance. According to the used basic network, FaceNet has multiple implements, and the adopted FaceNet in our work is based on Inception-Resnet-V1 network [40] illustrated in Fig. 3(a). The training set is MS-Celeb-1 M [12] and the dimension of output embedding is 128. In our work, the output feature of Reduction-B block, dropout layer and the final embedding within FaceNet are used to construct training losses for ApaNet.

ApaNet Inspired by Generative Adversarial Networks [10], we use eight residual blocks and the layers within each residual block to construct a generative network, which are illustrated in Fig. 3(b). Except for the last convolutional layer, the sizes of all convolutional filters in the network are $3 \times 3 \times 64$ and $9 \times 9 \times 3$ for the last convolutional layer. In addition, Batch Normalization (BN) layers is added to normalize the input (to have zero mean and variance), which is beneficial for stability training. Considering the bounded activation allows the model to learn more quickly to saturate and cover the color space of the training distribution, the ReLU activation is used in the ApaNet. Meanwhile, we use Tanh activation function in the output layer of ApaNet to achieve its rapid convergence. Since we more concern the classification result of the output image than its visual perceptual quality, we do not adopt the "discriminator" part as generator supervision. Instead, we attempt to seek a loss function directly relating to verification performance and induce more semantic features for the output image. Overall, the learning of ApaNet is efficient and stable, and as well as has the direct

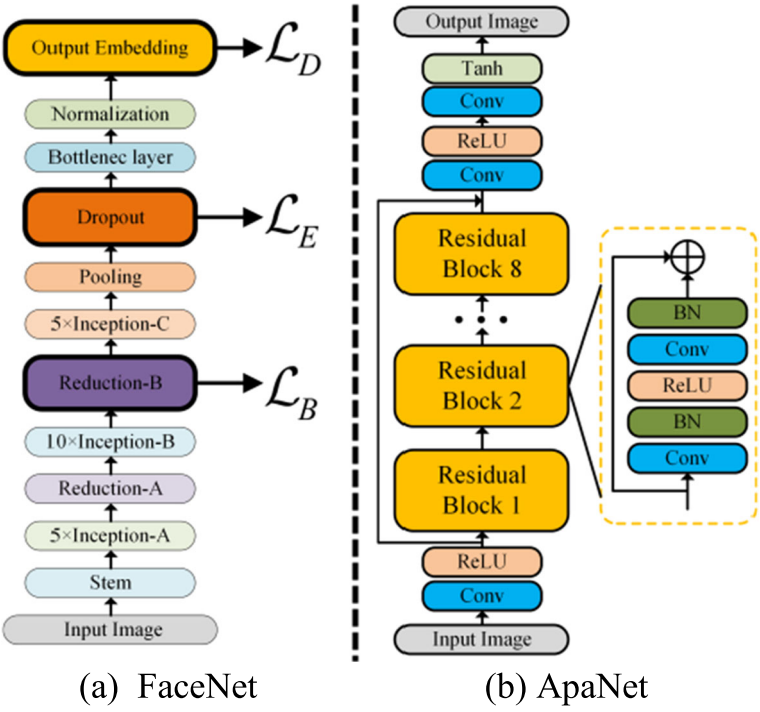

Fig. 3 The structure of FaceNet (**a**) and ApaNet (**b**)

connection with the performance of the target network. The optimization is discussed in detailed in section 3.3.

3.3 The optimization of ApaNet

Since we more concern the verification result of the output image than its visual perceptual quality, we leverage the representations extracted from middle and high layers of FaceNet due to their more semantic concepts rather than the difference between pixel values. In other words, for each pair of output image and legitimate image, minimizing loss function should encourage them to be close in middle and high layer feature space so as to recover the genuine identity of the adversarial example as far as possible. Fortunately, the optimization target coincides with the visual perceptual quality which is demonstrated by our experiment results (see Fig. 4). With an increasing of the number of selected feature maps, the computing consumption will be more intense whereas loss will more precisely represent the discrepancy between output and legitimate image, which implies an efficiency and efficacy trade-off. Thus, we extract **three** features from Reduction-B block output, dropout layer output and the final 128-dimensional embedding output as descriptors for the output image and legitimate image respectively. Then we calculate respective distance between the two descriptors, and take the weighted sum of the three distances as loss function. In following detailed explanation I_O and I_L denote output image during training and legitimate image respectively.

The first loss item The outputs of Reduction-B block are the aggregations of the features extracted from the previous blocks. $\phi_B(x)$ denotes the feature extracted from the output of Reduction-B block and has a shape $C_B \times H_B \times W_B$. This loss item $\mathcal{L}_B$ is defined as follows:

$$\mathcal{L}_B = \frac{1}{C_B H_B W_B} \|\phi_B(I_O) - \phi_B(I_L)\|_2^2 \tag{1}$$

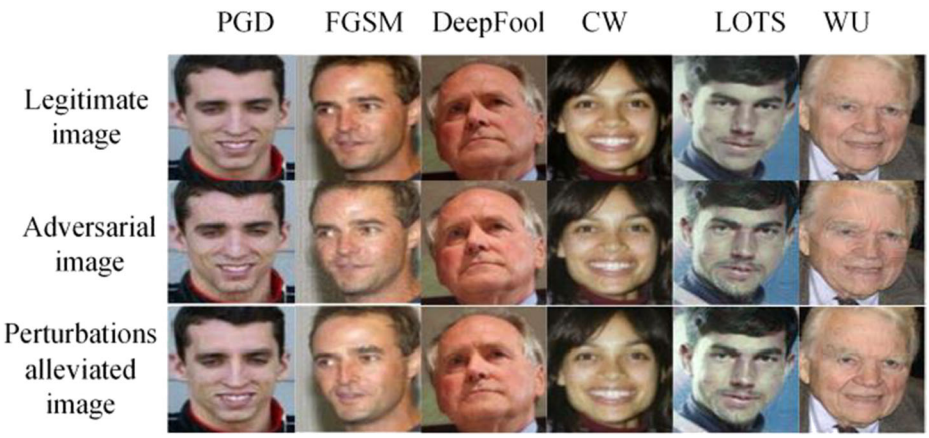

Fig. 4 A set of images from LFW dataset that including legitimate images, adversarial images and corresponding perturbations alleviated images by ApaNet for different white-box attacks

The second loss item Dropout is a simple way to prevent neural networks from over-fitting and improves the performance of neural networks on supervised learning tasks. The feature maps of this layer in FaceNet plays a very important role in semantic representation. $\phi_D(x)$ denotes the feature extracted from the output of dropout layer and has a shape $C_D \times H_D \times W_D$. This loss item $\mathcal{L}^D$ is calculated as follows:

$$\mathcal{L}_D = \frac{1}{C_D H_D W_D} \|\phi_D(I_O) - \phi_D(I_L)\|_2^2 \tag{2}$$

The third loss item The loss item $\mathcal{L}^E$ is to measure verification task errors for I_O and I_L. It depends on their embeddings of FaceNet and the squared Euclidean distance d between them. This loss item $\mathcal{L}_B$ is defined as follows:

$$d = \sum_{k=1}^{128} \left(\boldsymbol{E}_k^{I_O} - \boldsymbol{E}_k^{I_L}\right)^2 \tag{3}$$

$$Score = \begin{cases} 0.5 + \dfrac{(d-\eta) \times 0.5}{4-\eta}, & d > \eta \\ \dfrac{0.5 \times d}{\eta}, & d < \eta \end{cases} \tag{4}$$

$$\mathcal{L}_E = -log(1-score) \tag{5}$$

where η is a threshold, $\boldsymbol{E}^{I_O} \in R$ and $\boldsymbol{E}^{I_L} \in R$ respectively denote the embedding for I_O and I_L. The definition of this loss item implies that the verification result for output image and legitimate image is expected to be the same identity that is towards the ultimate verification goal.

In sum, the final joint loss function $\mathcal{L}_{Feature}$ is formed with a weighted sum of the three loss items:

$$\mathcal{L}_{Feature} = \alpha \cdot \mathcal{L}_B + \beta \cdot \mathcal{L}_B + \gamma \cdot \mathcal{L}_B \tag{6}$$

Considering the magnitude difference above the items, the weighting values α, β, γ are properly set as 1, 100 and 0.1 to adjust and normalize the value distribution range of the three losses items.

4 Experiment

4.1 Datasets

We evaluate our method on three datasets, including LFW, YouTube Faces DB and CASIA-FaceV5. All datasets are processed by MTCNN [49] for face detection and cropping.

LFW Labeled Faces in the Wild (LFW) is an academic dataset for face authentication which contains more than 13,000 face images of 5749 people.

YouTube faces DB It is a well-known dataset that has been widely used in the field of face recognition. Its organization is similar to LFW and pairs of video frame sequences are constructed instead of pairs of images in LFW.

CASIA-FaceV5 It is an Asian face dataset collected by Chinese Academy of Sciences which contains 2500 colour facial images of 500 subjects.

4.2 Experimental setting

We evaluate the defense performance of the proposed ApaNet on face verification task. The training set used to learn ApaNet is generated from LFW dataset, and validation set is constructed from the three datasets.

Evaluate rules: given a pair of two face images, a squared L_2 distance threshold is used to determine the classification of same and different. Specifically, two embedding are extracted by FaceNet for an input image and a reference image respectively and then the distance between them is calculated using Eq.(3). Compared with the threshold, if the distance is larger, the input image is verified as an identity different from the reference image; otherwise verified as the same identity as the reference image. The performance of ApaNet is evaluated by a ratio of number of correctly verified image pair to total number of image pairs. In following, we will use 'accuracy' to refer to the performance.

Training set First, we choose 2000 pairs of same-identity images and 2000 pairs of different-identity images from LFW dataset. Then we use PGD in APPENDIX A.1 to attack the one in the pair of same-identity images and generate dodging adversarial examples. Still attack the one in the pair of different-identity images to generate impersonation adversarial examples. These adversarial examples and corresponding legitimate ones compose a training set to learn the perturbations alleviation network. Required parameters are: the attack strength $\varepsilon = 0.1$, the attack step size $\alpha = 0.01$, the number of attack iterations $n = 20$.

Validation set with threshold According to [33], we choose approximately equal numbers of pairs of same-identity and different-identity images in each dataset and calculate their distances. The optimal distance threshold is selected under the equal error rate assessment of same-identity verification and different-identity verification. Table 2 shows the number of image pairs to calculate threshold on the LFW, YouTube Faces DB and CASIA-FaceV5 datasets which are also used as validation sets to evaluate ApaNet. In detail, the pairs of same-identity images are used to evaluate dodging attack, and the pairs of different-identity images are used to evaluate impersonation attack. The training set and validation set are completely independent in terms of image and identity.

Table 2 The number of images for validation and the corresponding thresholds on each dataset

	LFW	YouTube Faces DB	CASIA-FaceV5
Same-identity	1135	333	740
Different-identity	1195	333	880
Threshold	1.10	0.95	0.48

4.3 Experiment results

4.3.1 The defending against white-box attacks

We evaluated the effectiveness of ApaNet on LFW, YouTube Faces DB and CASIA-FaceV5 datasets. In this experiment, we compare the proposed defense with other perturbation cleaning methods including Randomization [23], ComDefend [45], TVM [13] and Gaussian blurring [8]. We evaluate their performances on six white-box attacks: PGD [22], FGSM [9], DeepFool [25], CW [37], LOTS [32] and WU[1] in Table 3. We leave the detailed attack algorithms in APPENDIX A. Each testing set contains the dodging and impersonation adversarial examples. We report the performance of ApaNet under three datasets in Tables 3, 4 and 5 respectively. For YouTube Faces DB and CASIA-FaceV5 datasets, we just choose Gaussian blurring with the best defense performance as the comparison method. It is obvious that ApaNet has the best defense performance in FaceNet compared to the comparison methods. In the LFW dataset, FaceNet could hardly correctly identify adversarial examples. However, under the protection of ApaNet, the recognition accuracy of all kinds of adversarial examples in FaceNet reached more than 95%. In the YouTube Faces DB and CASIA-FaceV5 datasets, the recognition accuracies of all kinds of adversarial example in FaceNet are above 90% and 75% under our defense. The results indicate that ApaNet learned with PGD examples has a satisfied generalization for different attacks and its joint use with FaceNet performs best among the compared approaches. For the evaluated attacks, a collection of legitimate images, adversarial images and perturbations alleviated images by ApaNet are shown in Fig. 4. We illustrate more visual results in APPENDIX B Figs. 5, 6 and 7.

4.3.2 The defending against black-box attacks

In this experiment, we evaluate the ability of ApaNet on defending against black-box attacks on LFW, YouTube Faces DB and CASIA-FaceV5 datasets. Here, the transfer-based attacks [29] and the NATTACK [19] are selected to verify its effectiveness. For the transfer-based attacks, we choose CosFace model as an alternative model due to its availability and its fine performance for face verification. Then we attack this model using the above white-box attack method to generate adversarial examples for FaceNet. It needs to be explained that LOTS and WU which designed for FaceNet don't support transfer-based attacks. On the other hand, NATTACK is a excellent black-box attack method which can defeat both vanilla DNNs and various defence techniques developed recently. Compared to white-box attacks, black-box attacks have a lower success rate against FaceNet without any defense, but they are more of a threat to real-world facial-recognition systems. The results illustrated in Table. 6 confirm that ApaNet learned using PGD examples also has a flexible adaptation for transfer-based black-box attacks.

4.3.3 The defending against defense-aware attacks

We assume that the adversary knows our proposed defence in advance and also has knowledge of the perturbations alleviation network (defense-aware attack). In this experiment, we test the attacking ability of PGD when it is used to attack the two networks simultaneously, namely

[1] https://github.com/ppwwyyxx/Adversarial-Face-Attack

Table 3 The accuracy of FaceNet under different white-box attacks (Dodging /Impersonation) on LFW dataset (%)

LFW	PGD	FGSM	DeepFool	CW	LOTS	WU
No Defense	0.1/7.7	8.8/17.1	1.3/7.1	0.3/7.5	3.0/6.4	2.3/2.1
Randomization	55.8/61.8	65.3/61.8	74.5/76.3	72.1/71.5	8.0/11.8	7.5/8.5
ComDefend	3.6/5.4	31.8/64.9	9.3/7.0	3.6/7.0	24.5/32.5	21.0/28.0
TVM	8.1/4.4	46.0/47.9	6.6/7.6	5.3/4.2	15.2/17.8	29.4/35.0
Gaussian blurring	78.4/78.6	84.8/85.9	86.5/86.1	86.1/86.6	12.0/14.5	15.5/16.6
ApaNet	**99.2/98.3**	**96.9/99.4**	**98.9/98.6**	**98.8/99.9**	**70.0/83.6**	**96.3/100.0**

The bold entries in table means our proposed method and the best results

Table 4 The accuracy of FaceNet under different white-box attacks (Dodging /Impersonation) on YouTube Faces DB dataset (%)

YouTube Faces DB	PGD	FGSM	DeepFool	CW	LOTS	WU
No Defense	11.0/32.5	13/33.0	7.4/19.7	26.9/37.6	23.2/53.9	17.5/34.9
Gaussian blurring	72.5/74.6	76.5/77.1	77.8/76.4	74.5/77.8	56.5/63.4	66.7/68.6
ApaNet	**92.5/93.6**	**94.3/95.6**	**91.9/92.5**	**94.8/93.0**	**97.5/98.2**	**98.7/99.5**

The bold entries in table means our proposed method and the best results

perturbations alleviation network and FaceNet. For comparison, we also list the attacking rate of sole FaceNet attacked. The results in Table 7 confirm that the using of perturbations alleviation network has greatly boosted the ability of counteracting attacks.

4.3.4 The ablation experiment on different loss functions

This experiment aims to confirm the necessity and effectiveness of joint loss function and compare the performance using different sole loss item. And the dodging and impersonation adversarial examples are generated by PGD on LFW datasets. The results in Table 5 indicate that joint using of the three loss items outperform the other sole loss item with a large margin, which is consistent with our expectations. Especially, the result using the difference between pixel values as loss function is the lowest for dodging and impersonation adversarial examples and it has confirmed that essentially different characteristics between output image and legitimate image are conveyed by the middle and high layers. As the legitimate image is also processed by ApaNet during testing, we also test the performance on the legitimate examples to explore how ApaNet impacts them. The last row of Table 8 shows there is only a slight

Table 5 The accuracy of FaceNet under different white-box attacks (Dodging /Impersonation) on CASIA-FaceV5 dataset (%)

CASIA-FaceV5	PGD	FGSM	DeepFool	CW	LOTS	WU
No Defense	12.0/23.5	11.5/26.8	17.4/27.2	16.5/28.7	19.5/23	2.2/7.3
Gaussian blurring	47.8/57.2	52.3/56.5	47.5/48.9	41.1/48.6	47.5/49..7	65.2/43.2
ApaNet	**83.4/70.2**	**87.3/77.9**	**75.3/76.5**	**76.3/75.8**	**75.8/73.3**	**78.9/61.0**

The bold entries in table means our proposed method and the best results

Table 6 The accuracy of FaceNet under different black-box attacks (Dodging/Impersonation) on different datasets (%)

Dataset	Defense Type	PGD	FGSM	DeepFool	CW	NATTACK
LFW	No Defense	62.6/15.5	70.9/12.5	76.5/10.5	62.3/15.1	14.3/16.6
	ApaNet	78.3/92.6	79.3/92.3	80.1/91.5	77.4/92.7	93.2/98.6
YouTube Faces DB	No Defense	54.2/23.5	65.3/34.5	57.2/25.6	53.2/19.8	7.5/14.2
	ApaNet	73.7/87.5	82.5/97.6	77.7/86.7	74.9/84.3	89.5/94.3
CASIA-FaceV5	No Defense	57.5/25.0	75/27.4	53.4/19.5	49.5/14.1	10.7/13.6
	ApaNet	70.8/67.8	75.9/75.7	69.4/74.4	72.3/73.4	87.5/85.3

Table 7 The success rates of PGD attacking FaceNet and the combination of ApaNet and FaceNet (%)

Dataset	Targeted model	Dodging	Impersonation
LFW	FaceNet	99.9	92.3
	ApaNet+FaceNet	35.1	38.4
YouTube Faces DB	FaceNet	89.0	67.5
	ApaNet+FaceNet	30.0	23.5
CASIA-FaceV5	FaceNet	88	76.5
	ApaNet+FaceNet	36.7	40.5

decrease for accuracy. In sum, ApaNet optimized with the joint loss function is considerably effective on diminishing adversarial perturbations.

4.4 Discussion

According to the Manifold Hypothesis [38], for most AI tasks, the full sample space is located in high dimensions, but the effective data we can grasp lie actually on a manifold with a lower dimension. This suggests that the legitimate examples are on a manifold, and adversarial examples are off the manifold with high probability. From the flexible adaptability of ApaNet to various white-box attacks and black-box attacks, we can deduce that the adversarial examples produced by PGD attack locate in a dense region of adversarial examples and can be taken as an anchor in that region. In addition, according to the perspective that the PGD attacked examples leave the manifold, we infer that they are not far away from manifold as the legitimate examples on manifold have a slight decline of accuracy after the legitimate examples are processed by ApaNet.

Table 8 The accuracy of FaceNet under the protection of ApaNet optimized with different loss functions (%)

Input type	Pixel-level loss	Reduction loss	Dropout loss	Embedding loss	Joint loss
Dodging	10.9	51.6	70.9	72.6	**99.2**
Impersonation	52.4	73.8	85.6	87.4	**98.3**
Legitimate	98.1	97.2	96.3	95.7	**97.8**

The bold entries in table means our proposed method and the best results

5 Conclusion

In this paper, we investigate how to defend target DNNs in face verification scenario by alleviating adversarial perturbations injected into the input facial image. Specifically, we design ApaNet which is implemented with stacked residual structures. Then we employ FaceNet as target network and PGD attack to generate dodging and impersonation adversarial examples, along with the corresponding legitimate counterparts as supervision. To have a supervised learner for ApaNet, we define a joint loss function which measures the discrepancy between the output image and legitimate image depending on the representations from the output of Reduction-B block, dropout layer and the final embedding layer of FaceNet. The representations of these layers convey more semantic information and are crucial for alleviating effects. The ablation experiment confirms the advantage of joint using of the three loss items over the sole loss item. In addition to PGD attack, ApaNet has shown a satisfied generalization on FGSM, CW, DeepFool, LOTS, WU attacks and even is capable of resisting black-box attacks including transfer-based attacks and NATTACK. Especially, compared with several currently available defensive techniques, the proposed ApaNet performs better. It is worth emphasizing that the training of ApaNet is based on LFW dataset, its testing has extended to YouTube Faces DB and CASIA-FaceV5 so as to show its generalization across datasets.

Although we focus on face verification task, the mechanism proposed in our work can be readily extended to other applications, for instance, image classification, object detection and semantic segmentation. In addition, the network that serves for constructing loss function, like FaceNet, could be an adversarial trained version. In the future, we will develop investigation towards these aspects.

Appendices

A detailed attack algorithm

The method of attacking FaceNet through classifier conversion

Once the feature is extracted from embedding of FaceNet, Eq.(3 $\sim$ 5) can be used to convert attacking FaceNet into attacking a binary classifier in our work. At this point, Eq.(5) is used as the loss function of classifier for implementing attack based on gradient. Take FGSM and PGD for example. The adversarial examples can be expressed in the following formula ('+'for dodging, '-'for impersonation):

$$X_{FGSM}^{adv} = X \pm \varepsilon \cdot sign\left(\nabla_X \mathcal{L}_E\left(\theta, X, X_{refer}\right)\right) \tag{7}$$

$$X_{PGD}^{adv} = Clip_{X,\varepsilon}\left[X_k^{adv} \pm \alpha sign\left(\nabla_X \mathcal{L}_E\left(\theta, X, X_{refer}\right)\right]\right. \tag{8}$$

Note that in our experiments, Eq.(5) is also used for the optimization of DeepFool and CW attacks.

The method of attacking FaceNet through LOTS

Layerwise origin-target synthesis (LOTS) generates adversarial examples by imitating the deep features of the target. Here, LOTS use Euclidean distance to measure the input itself or the discrepancy between the adversarial input. To be more effective, we attack the activation of Reduction-B block of FaceNet, which is also one of the layers used for training. The targeted (Impersonation) and untargeted (Dodging) adversarial examples can be expressed in the following formula:

$$X_{LOTS}^{target} = Clip_{X,\varepsilon}\left[X_k^{adv} - \alpha\nabla_X \left\| \phi_B\left(X_k^{adv}\right) - \phi_B\left(\mathrm{X}_{target}\right)\right\|_2\right] \tag{9}$$

$$X_{LOTS}^{untargeted} = \mathrm{C}lip_{X,\varepsilon}\left[X_k^{adv} + \alpha\nabla_X \left\| \left(X_k^{adv}\right)\right\|_2\right] \tag{10}$$

The method of attacking FaceNet through WU

The method proposed by Dr. Yuxin Wu in 2018 GeekPwn CAAD is a attack against FaceNet. First, N face images of target identities are collected, and N embedding vectors $\boldsymbol{V_i}$ are extracted by running FaceNet. Then adversarial examples are generated by minimizing the average distance of input image and embedding of N images:

$$minimize\ L = \frac{1}{N}\sum_{i=1}^{N} dist(\boldsymbol{E}_x, V_i) \tag{11}$$

The adversarial examples generation method refers to PGD and can be expressed in the following formula:

$$X_{WU,k+1}^{adv} = Clip_{X,\varepsilon}\left[X_k^{adv} \pm \alpha \cdot \nabla_X\left(\frac{1}{128}\boldsymbol{E}\left(X_k^{adv}\right) \times \boldsymbol{E}^{\boldsymbol{T}}\left(X_{target}\right)\right)\right] \tag{12}$$

In our experiment, required parameters for WU attack are set as: the attack strength ε=8, the attack step size α=0.9 and the number of attack iterations $k = 200$.

Qualitative examples

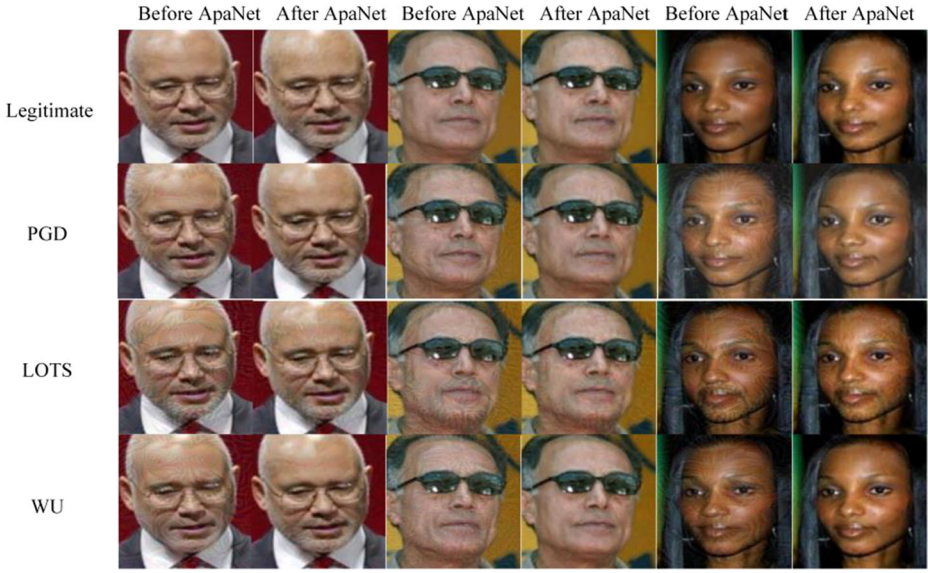

Fig. 5 A collection of legitimate images, adversarial images (PGD, LOTS and WU attacks) and perturbations alleviated images after ApaNet in the LFW dataset

Fig. 6 A collection of legitimate images, adversarial images (PGD, LOTS and WU attacks) and perturbations alleviated images after ApaNet in the YouTube Faces DB dataset

Fig. 7 A collection of legitimate images, adversarial images (PGD, LOTS and WU attacks) and perturbations alleviated images after ApaNet in the CASIA-FaceV5 dataset

Acknowledgements This work was supported by Natural Science Foundation of Shanghai under Grant No. 20ZR1419900 and Shanghai Committee of Science and Technology under Grant (No.21511102605).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Becerra-Riera F, Morales-González A, Méndez-Vázquez H (2019) A survey on facial soft biometrics for video surveillance and forensic applications. Artif Intell Rev 52(2):1155–1187
2. Boutros F, Siebke P, Klemt M, Damer N, Kirchbuchner F, Kuijper A (2021) Pocketnet: extreme lightweight face recognition network using neural architecture search and multi-step knowledge distillation. arXiv preprint arXiv:2108.10710
3. Chhabra S, Singh R, Vatsa M, Gupta G (2018) Anonymizing k-facial attributes via adversarial perturbations. arXiv preprint arXiv:1805.09380
4. 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), pp 1979–1988
5. Deb D, Liu X, Jain AK (2020) Faceguard: a self-supervised defense against adversarial face images. arXiv preprint arXiv:2011.14218
6. Duan R, Ma X, Wang Y, Bailey J, Qin A K, Yang Y (2020) Adversarial camouflage: hiding physical-world attacks with natural styles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1000–1008
7. Fan W, Sun G, Su Y, Liu Z, Lu X (2019) Integration of statistical detector and gaussian noise injection detector for adversarial example detection in deep neural networks. Multimed Tools Appl 78(14):20409–20429

8. 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), pp 1-7
9. Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572
10. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Adv Neural Inf Proces Syst 27
11. 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
12. Guo Y, Zhang L, Hu Y, He X, Gao J (2016) Ms-celeb-1m: a dataset and benchmark for large-scale face recognition. In: European conference on computer vision (ECCV), pp 87–102
13. Guo C, Rana M, Cisse M, Van Der Maaten L (2017) Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117
14. Hu J, Liao X, Wang W, Qin Z (2021) Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. IEEE Trans Circuits Syst Video Technol 32:1089–1102
15. Huang G B, 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, pp
16. Jia X, Wei X, Cao X, Foroosh H (2019) Comdefend: an efficient image compression model to defend adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6084–6092
17. Kumar A, Singh N, Kumar P, Vijayvergia A, Kumar K (2017) A novel superpixel based color spatial feature for salient object detection. In: 2017 conference on information and communication technology (CICT), pp 1-5
18. Kumar K, Kumar A, Bahuguna A (2017) D-cad: deep and crowded anomaly detection. In: Proceedings of the 7th international conference on computer and communication technology, pp 100-105
19. Li Y, Li L, Wang L, Zhang T, Gong B (2019) Nattack: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In: International Conference on Machine Learning, pp. 3866–3876
20. Liao X, Yin J, Chen M, Qin Z (2020) Adaptive payload distribution in multiple images steganography based on image texture features. IEEE Trans Dependable Secure Comput
21. Liao X, Li K, Zhu X, Liu KR (2020) Robust detection of image operator chain with two-stream convolutional neural network. IEEE J Sel Top Signal Process 14(5):955–968
22. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2017) Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083
23. Massoli FV, Carrara F, Amato G, Falchi F (2021) Detection of face recognition adversarial attacks. Comput Vis Image Underst 202:103103
24. Mirjalili V, Ross A (2017) Soft biometric privacy: retaining biometric utility of face images while perturbing gender. In: 2017 IEEE international joint conference on biometrics (IJCB), pp 564-573
25. Moosavi-Dezfooli S-M, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574–2582
26. Negi A, Chauhan P, Kumar K, Rajput R (2020) Face mask detection classifier and model pruning with keras-surgeon. In: 2020 5th IEEE international conference on recent advances and innovations in engineering (ICRAIE), pp 1-6
27. Negi A, Kumar K, Chaudhari N S, Singh N, Chauhan P (2021) Predictive analytics for recognizing human activities using residual network and fine-tuning. In: International Conference on Big Data Analytics, pp. 296–310
28. Papernot N, McDaniel P, Wu X, Jha S, Swami A (2016) Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE symposium on security and privacy (SP), pp 582-597
29. Papernot N, McDaniel P, Goodfellow I (2016) Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277
30. Ren K, Zheng T, Qin Z, Liu X (2020) Adversarial attacks and defenses in deep learning. Engineering 6(3): 346–360
31. Rozsa A, Günther M, Rudd E M, Boult T E (2016) Are facial attributes adversarially robust? In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp 3121–3127
32. Rozsa A, Günther M, Boult TE (2017) Lots about attacking deep features. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp 168–176

33. 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.
34. 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
35. Sharma S, Kumar K (2021) Asl-3dcnn: American sign language recognition technique using 3-d convolutional neural networks. Multimed Tools Appl 80(17):26319–26331
36. Sharma S, Kumar K, Singh N (2017) D-Fes: deep facial expression recognition system. In: 2017 conference on information and communication technology (CICT), pp 1-6
37. Sriram S, Simran K, Vinayakumar R, Akarsh S, Soman K (2019) Towards evaluating the robustness of deep intrusion detection models in adversarial environment. In: International Symposium on Security in Computing and Communication, pp. 111–120
38. Stutz D, Hein M, Schiele B (2019) Disentangling adversarial robustness and generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6976–6987
39. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199
40. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence, pp
41. Taigman Y, Yang M, Ranzato MA, Wolf L (2014) Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1701-1708
42. Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P (2017) Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204
43. Wang H, Wang Y, Zhou Z, Ji X, Gong D, Zhou J, Li Z, Liu W (2018) Cosface: large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5265–5274
44. Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity. In: CVPR 2011, pp. 529–534
45. Xie C, Wang J, Zhang Z, Ren Z, Yuille A (2017) Mitigating adversarial effects through randomization. arXiv preprint arXiv:1711.01991
46. Xie C, Tan M, Gong B, Yuille A, Le Q V (2020) Smooth adversarial training. arXiv preprint arXiv: 2006.14536
47. Yi D, Lei Z, Liao S, Li SZ (2014) Learning face representation from scratch. arXiv preprint arXiv: 1411.7923
48. Yuan X, He P, Zhu Q, Li X (2019) Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst 30(9):2805–2824
49. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503