Crafting Transferable Adversarial Examples Against Face Recognition via Gradient Eroding

Huipeng Zhou, Yajie Wang, Yu-an Tan, Shangbo Wu, Yuhang Zhao, Quanxin Zhang, and Yuanzhang Li

***Abstract*—In recent years, deep neural networks (DNNs) have made significant progress on face recognition (FR). However, DNNs have been found to be vulnerable to adversarial examples, leading to fatal consequences in real-world applications. This article focuses on improving the transferability of adversarial examples against FR models. We propose gradient eroding (GE) to make the gradient of the residual blocks more diverse, by eroding the back-propagation dynamically. We also propose a novel black-box adversarial attack named corrasion attack based on GE. Extensive experiments demonstrate that our approach can effectively improve the transferability of adversarial attacks against FR models. Our approach overperforms 29.35% in fooling rate than state-of-the-art black-box attacks. Leveraging adversarial training with adversarial examples generated by us, the robustness of models can be improved by up to 43.2%. Besides, corrasion attack successfully breaks two online FR systems, achieving a highest fooling rate of 89.8%.**

***Impact Statement*—While DNNs are widely applied in face recognition tasks, they are susceptible to adversarial attacks even under black-box scenarios. Black-box adversarial examples are generally crafted against an ensemble of virtual models for the sake of transferability, where these virtual models are generated by eroding intermediate structures of a base model. However, this erosion mechanism impairs the virtual model's accuracy, which in turn, causes miscalculations for the adversarial loss, damaging the attack's effectiveness. We solve this problem by introducing Gradient Erosion (GE) so that the crafted model ensemble remains diverse while not losing its accuracy, thereby improving the effectiveness of adversarial attacks against black-box face recognition systems impressively. Our proposed approach implies urgent improvements lying within face recognition systems deployed in both lab experimental environments and production commercial products.**

***Index Terms*—Adversarial example, black-box attack, face recognition (FR), transfer attack, transferability.**

I. Introduction

IN RECENT years, deep neural networks (DNNs) have made significant progress in face recognition (FR) [1], [2], [3], [4], [5], [6], [7]. FR is a recognition technology for authentication that has been widely deployed in applications in various fields such as financial payments, security, and the military. DNN-based FR systems can easily identify people by analyzing their facial features, with performances far exceeding traditional algorithms. However, security-sensitive scenarios call for not only recognition efficiency, but also robustness against potential malicious attacks. Hence, the malicious effects of potential attacks on the system are to be considered. Researchers have found that DNN-based systems are highly susceptible to adversarial attacks, leading to misjudgments in FR [8], [9] systems with nasty or even fatal consequences. Therefore, researching the generation process of adversarial examples to understand the intrinsic mechanism and developing new attack methods can help facilitate the game between adversarial attacks and defenses. Similar to adversarial attacks [10], [11], [12], [13], [14], more and more security problems in new scenarios continue to rise, such as backdoor attacks [15] and privacy protection [16], [17], [18].

Adversarial examples [19], [20], [21] are malicious images with imperceptible perturbations that cause DNNs to make mispredictions with high confidence [22], [23], [24]. Attackers typically use transfer-based methods to attack black-box models, most commonly by attacking local substitution models to generate adversarial examples and exploiting the transferability of the adversarial examples [20] to attack black-box models. Recent works have found that attacks on ensembles of substitution models can improve the transferability of adversarial examples. Researchers have further developed attack methods based on ensemble models [25] to enhance transferability. The momentum iterative attack (MI-FGSM) [26] is designed to stabilize the update direction by integrating momentum terms to avoiding local optima. The diversity input attack (DI-FGSM) [27] randomly resizes and fills the input with a fixed probability at each iteration. The scale-invariant method (SI-FGSM) [28] optimizes the ingestion of the scaled image ensemble at each iteration. However, these methods are proposed in the context of image classification tasks. Therefore, we mainly focus on improving the transferability of adversarial examples in black-box scenarios for FR tasks.

Leveraging the transferability of adversarial examples to attack black-box models [20], many solutions have been proposed, such as attacking substitution models [29] or an ensemble of substitution models [25], [26], [30], [31]. Researches show that the transferability of adversarial examples is positively correlated

Manuscript received 3 October 2022; revised 5 December 2022; accepted 25 February 2023. Date of publication 6 March 2023; date of current version 8 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant U1936218, and Grant 62072037. This article was recommended for publication by Associate Editor Marley Vellasco upon evaluation of the reviewers' comments. *(Corresponding author: Yajie Wang.)*

Huipeng Zhou, Quanxin Zhang, and Yuanzhang Li are with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: zhouhuipeng@bit.edu.cn; zhangqx@bit.edu.cn; popular@bit.edu.cn).

Yajie Wang, Yu-an Tan, Shangbo Wu, and Yuhang Zhao are with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: wangyajie0312@foxmail.com; tan2008@bit.edu.cn; spencer.wushangbo@gmail.com; zhaoyuhang@bit.edu.cn).

Digital Object Identifier 10.1109/TAI.2023.3253083

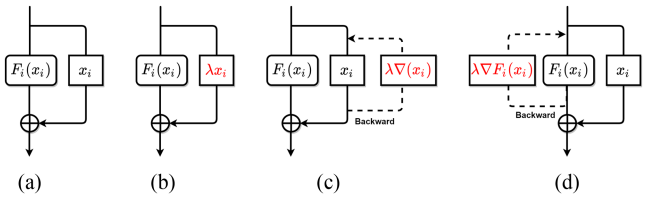


Fig. 1. Erosion is performed for residual networks (a) for standard residual connections, (b) for Ghost erosion mechanism to erode skip connections, (c) for skip connections when they are back-propagated, and (d) for residual blocks when they are back-propagated. The erosion method in (b) loses the forward propagation accuracy of the model, while our proposed (c) and (d) do not affect the model accuracy and also improve the transferability of adversarial attacks.

with the diversity of ensemble models. Li et al. proposed the Ghost mechanism to erode deep residual neural networks [32] to improve the transferability of adversarial examples, as shown in Fig. 1(b). The basic principle of their work is to generate virtual models on a base network (a network trained from scratch). These virtual models are generated by dynamic erosion on the intermediate structure of the base network. However, this erosion mechanism loses accuracy when propagating forward, leading to a large discrepancy between predictions and ground truth. Such discrepancy results in a huge difference in the calculation of the adversarial loss. To solve this problem, we propose gradient eroding (GE) of the residual structure, which can enrich the diversity of ensemble models without losing accuracy in the forward propagation.

We propose two attack schemes. First, in back-propagation, we dynamically erode the gradient of the skip-connected part of the residual structure, as in Fig. 1(c). Second, we erode gradients dynamically, which propagates the residual block part of the residual structure gradient in back-propagation, as in Fig. 1(d). The experimental results demonstrate that this dynamic erosion gradient approach in Fig. 1(d) can effectively improve transferability up to 14.8% for black-box attacks compared with existing methods. More critically, eroding only the back-propagation gradient ensures that our model accuracy is not sacrificed. Based on our proposed GE, we propose a novel black-box adversarial attack, called corrasion attack, which can achieve a fooling rate of 19.68% higher than other black-box adversarial attack. We even achieve a fooling rate of 29.35% higher than other aggressive approaches. Moreover, adversarial training using our proposed corrasion attack method can help improve the robustness of DNNs.

Extensive experiments demonstrate our proposed corrasion attack's power and the effectiveness of GE. We use 2000 pairs of face images in our local library of 15 FR models to perform our black-box adversarial attack. Results show that the adversarial examples generated by corrasion attack have higher transferability and black-box fooling rate. Also, combining proposed dynamic erosion back-propagation gradient mechanism with other adversarial attack methods can significantly enhance black-box attacks' fooling rate. Moreover, we use corrasion attack to train two FR models against each other, and we improve the robustness of the models by 43.2%. Finally, we also performed an attack against two online FR systems and achieved a high fooling rate.

In summary, the main contributions of this article are as follows.

1) We propose GE, a dynamic back-propagation gradient erosion mechanism for residual blocks in residual network structures. GE can produce more diverse gradient information, which can effectively improve the transferability of adversarial examples.
2) Based on GE, we proposed corrasion attack, a novel adversarial attack against black-box FR models. Corrasion attack can achieve a 19.68% higher fooling rate than other commonly used adversarial attacks. We attack two online FR systems and gain a fooling rate of 84.5% on AWS and 89.8% on Clarifai.
3) We use the proposed strategy for adversarial training on two FR models. The robustness of the models improve by 43.2% under five adversarial attack methods.

II. Related Work

A. Adversarial Attack

a) Fast Gradient Sign Method (FGSM): Goodfellow et al. proposed the FGSM [19]. The adversarial example x_{adv} is generated by maximizing loss function $J(x_{\text{adv}}, y^{\text{true}})$ with a one-step update, the update rule is

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y^{\text{true}})) \tag{1}$$

where x is the clean image, y^{true} is the ground-truth label for x, x_{adv} is the adversarial example, and ϵ is a l_∞ norms constraint.

b) Momentum Iterative Fast Gradient (MI-FGSM): Dong et al. proposed MI-FGSM [26]. By integrating momentum into the iterative process, a higher degree of transferability is brought to adversarial examples. The update rule is

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J\left(x_t^{\text{adv}}, y^{\text{true}}\right)}{\| \nabla_x J\left(x_t^{\text{adv}}, y^{\text{true}}\right) \|_1} \tag{2}$$

$$x_{t+1}^{\text{adv}} = \text{Clip}_x^{\epsilon}\left\{x_t^{\text{adv}} + \alpha \cdot \text{sign}(g_{t+1})\right\} \tag{3}$$

where g_t is the gradient of the image at step t and μ is the momentum.

B. Ghost Network

Ghost network [32] multiplies an erosion parameter by the skip-connected part of the residual network as in Fig. 1(b), and the erosion parameter is selected by sampling in the range of about 10% decrease in model accuracy to obtain a large number of local erosion models, which are generated by dynamically eroding some intermediate structures of the underlying network. Iterative attacks on adversarial examples can diversify the obtained image gradients and thus improve the transferability of adversarial examples and the fooling rate of attacking unknown black-box models. However, this mechanism has an obvious drawback in that it sacrifices the accuracy of the model. In other words, there is a large difference between the network output obtained by the Erode model and the final true network output obtained by the unErode model. This difference will have a large error when calculating the loss function back-propagation to solve the image gradient. We believe that the gradients thus solved can achieve gradient diversification. Still, it is at the

Algorithm 1: Corrasion Attack.

Input: Original human face image x, targeted face image x_t.
Parameter: Face feature extraction model F, the number of attack iterations is $iters$, number of sub-iterations m, the iteration step size is μ, the constraint ℓ_∞-norms is ϵ, random noise ξ is consistent with $\mathcal{N}(0, \delta^2 I)$, transformation function $T(x)$, kernel matrix W.
Output: Adversarial example x_{adv}.

$x_{\text{adv}} \leftarrow x$
$\text{output}_t \leftarrow F(x_t)$.
$G = 0$
while $i <$ iters **do**
  $G_i = 0$.
  while $j < m$ **do**
    $\text{output}_j \leftarrow F(T(S_j(x_{\text{adv}} + \xi_j)))$
    $J = \sqrt{\| F(x_{\text{adv}}) - F(x_t) \|_2^2}$
    $g_j = \frac{\nabla J(\text{output}_j, \text{output}_t)}{\|\nabla J(\text{output}_j, \text{output}_t)\|_1}$
    $G_i = G_i + g_j$
  end while
  $G_i = \frac{G_i}{m}$
  $G = \mu \cdot G + G_i$
  $x_{\text{adv}} = \text{clip}_{x,\epsilon}\{x_{\text{adv}} + \mu \cdot \text{sign}[W * (G)]\}$
end while
return x_{adv}

expense of model accuracy, and the obtained gradients do not completely describe the original gradient information of the adversarial examples. Therefore, we investigate this problem, and the details are described in Section III.

III. Methodology

A. Attack FR

Given an FR model $\mathbf{F}(x) : x \to R^d$ denotes the extraction of R^d normalized features of the input face image x, where $x \in \mathbf{X} \subset R^d$. The purpose of FR is to recognize as a specified face, that is, to calculate the distance between the normalized feature vectors output from a pair of face images, usually use cosine similarity to measure the distance between two normalized feature vectors. Given a pair of face images (x_a, x_b), where $x_a \in \mathbf{X}, x_b \in \mathbf{X}$. The distance between them is defined as

$$D(x_a, x_b) = \sqrt{\| F(x_a) - F(x_b) \|_2^2}. \tag{4}$$

Therefore, for the FR model, face verification process is to compare whether the distance $D(x_a, x_b)$ of a pair of faces is greater than a certain threshold δ. The final recognition result $S(x_a, x_b)$ is denoted as

$$S(x_a, x_b) = \begin{cases} 1 & D(x_a, x_b) \geq \delta \\ 0 & D(x_a, x_b) < \delta \end{cases}. \tag{5}$$

When the distance $D(x_a, x_b)$ is greater than or equal to the threshold δ, the recognition result is 1, which means that the two faces are the same person, and when the distance $D(x_a, x_b)$ is less than the threshold δ, the recognition result is 0, which means that the two faces are different people.

The attacker wishes to generate an adversarial example with the local white-box model to attack the FR model by using the transferability of the adversarial example. Given a pair of face images (x_s, x_t), the adversary wants to find an adversarial example x'_s satisfying $\| x'_s - x_s \|_\infty \leq \epsilon$ such that $S(x'_s, x_t) = 1$ — recognized as the same face. Thus, the problem is transformed into

$$\arg\min D(x_s, x_t), \quad \text{s.t.} \| x'_s - x_s \|_\infty \leq \epsilon. \tag{6}$$

B. Eroding Back Propagation of Residual Block

In the residual network, $F_i(\cdot)$ denotes the residual function in the residual block at the layer i, $F_i(x_i)$ denotes the output of the residual function at layer i, x_i denotes the skip-connected part at layer i, and λ denotes the erosion parameter. The residual structure in a standard residual network is shown in Fig. 1(a).

Ghost's erosion rule for residual networks multiplies the erosion parameter λ in its skip-connected part, as in Fig. 1(b), i.e., the ith layer skip-connected part changes to λx_i. However, the Ghost mechanism leads to a large discrepancy between the model's predictions and ground truth. Therefore, we want to make the obtained image gradients diverse without losing the model forward propagation accuracy. In light of this, we propose dynamic gradient erosion of the residual structure during network back-propagation to solve for the gradients, which will not affect the forward propagation of the model, and thus, the final output of the model. For the characteristics of the residual structure in the residual network, we propose two ways of gradient erosion in Fig. 1(c) and (d). Gradient erosion is performed for the skip-connection part and the residual block, respectively. Fig. 1(c) shows that the gradient is multiplied by the erosion parameter during the back-propagation of the skip-connected part at the ith level to make it $\lambda_i \nabla(x_i)$ as in (7). Fig. 1(d) represents the part of the residual block at the layer i given its gradient multiplied by the erosion parameter during its back-propagation so that it becomes $\lambda_i \nabla F_i(x_i)$ as in (8).

$$\nabla x_{i+1} = \lambda_i \nabla(x_i) + \nabla(F_i(x_i)) \tag{7}$$

$$\nabla x_{i+1} = \nabla(x_i) + \lambda_i \nabla(F_i(x_i)). \tag{8}$$

Like the Ghost erosion mechanism, our erosion parameters are dynamically generated to ensure the diversity of the obtained gradients. Our erosion parameters λ are a set, i.e., $\lambda = \{\lambda_1, \lambda_2, \lambda_3 \cdots \lambda_i\}$. For the ith residual layer, it corresponds to the erosion parameter λ_i, and each λ_i is obtained by random sampling from the range of (0, 1], which means that the set of λ obtained by each back-propagation is different. The erosion mechanism of the residual model using both ways in Fig. 1(c) and (d) can ensure that the accuracy of the model forward propagation is not lost and the image gradients acquired during back-propagation maintain the original diversity. We will conduct a detailed experimental discussion of both approaches in the experimental section.

C. Corrasion Attack

We use the momentum-based iterative approach as the baseline of our method. First, we use the commonly used technique of data augmentation [27] to transform the original face image x. A transformation function $T(\cdot)$ as in (9) is designed to translate and scale the adversarial example x_{adv} with random probability p before each round of iterative attack on the input. This enriches the diversity of the input samples.

$$T(x_{\text{adv}}, p) = T(x_{\text{adv}}), \quad \text{with probability } p. \tag{9}$$

Each round of iteration is subdivided into m rounds of subiterations. In the subiterations, we use the scale-invariant method [33]. Scale transformation of the image is performed as in (10), where $S_i(\cdot)$ denotes the scale transformation function. The image pixels are divided by 2 to the ith power in the m-round iterations, and the transformed adversarial examples are fed into the network model to solve for the gradient g. Finally, we take the average of the gradients obtained from the m-round iterations as in (11), where $J(\cdot)$ denotes the loss function generally chosen as the cosine similarity. The transformation of the adversarial examples with random scales effectively avoids the overfitting of the adversarial examples for a specific size due to the scale invariance of the deep learning model.

$$S_i\left(x_{\text{adv}}\right) = \frac{x_{\text{adv}}}{2^i}, \quad i \in [0, m]. \tag{10}$$

$$g = \frac{1}{m}\sum_{i=1}^{m} \nabla J(S_i(x_{\text{adv}})). \tag{11}$$

To reduce the variance of the gradient of the image acquisition, we add a certain range of random noise ξ in each round of iterations [28], as follows:

$$g = \frac{1}{m}\sum_{i=1}^{m} \nabla J(x_{\text{adv}} + \xi_i), \quad \xi_i \sim \mathcal{N}(0, \delta^2 I). \tag{12}$$

Finally, the image gradient is smoothed on the obtained image gradient using the kernel matrix W [34], as follows:

$$g = W * \nabla J(x)|_{x=x_{\text{adv}^*}}. \tag{13}$$

We use ℓ_∞-norms as a constraint on our method, G_t is the gradient accumulated in the tth round and y^{true} is the feature vector of our target attack face image. By combining (10)–(13), the update rule of our attack method is formulated as

$$x_{t+1}^{\text{adv}} = \text{clip}_{x,\epsilon}\left\{x_t^{\text{adv}} + \alpha \cdot \text{sign}\left[W * (G_t)\right]\right\} \tag{14}$$

$$\text{where} \quad G_t = \mu \cdot G_{t-1} + \frac{g_t}{\parallel g_t \parallel_1} \tag{15}$$

$$g = \frac{1}{m}\sum_{i=1}^{m} \nabla J(S_i(x_{\text{adv}}) + \xi_i, y^{\text{true}}) \quad \xi_i \sim \mathcal{N}(0, \delta^2 I). \tag{16}$$

The process of our approach is shown in Algorithm 1. The transferability of adversarial examples is further enhanced by adding gradient averaging, variance reduction, and data augmentation to the iterative attack.

IV. Experiments

A. Setup

Experiments are conducted using face feature extraction models from our local model library as white-box and black-box models. All models are sourced from the Internet, and we do not make any modifications. Our local model library contains 15 models are: InsightFace_mobilefacenet [35], InsightFace_IR_SE50 [35], CosFace [36], FaceNet_casia [37], FaceNet_vggface2 [37], ArcFace [35], MobileNet [38], ResNet50 [39], ShuffleNet [39], evoLVe_IR_152 [40], evoLVe_IR_50 [40], evoLVe_IR_50_Asia [40], MXNET_LResNet34E_IR [35], MXNET_LResNet50E_IR [35], and MXNET_LResNet100E_IR [35]. We focus our discussion on targeted attacks on FR models because we believe that targeted attacks in FR tasks can lead to greater harm and consequences.

In this article, we use 2000 pairs of face images from the Labeled Faces in the Wild (LFW) [41] dataset for attack-related experiments, using the CASIA-WebFace [42] dataset for adversarial training of the model. We select five currently used adversarial attack methods to compare with our approach, which are MI-FGSM [26], TI-FGSM [34], DI-FGSM [27], Vr-FGSM [28], and SI-FGSM [33] using cosine similarity as the loss function. In our method, the parameters are set to ϵ=20.4, iters=100, m=3, targeted attack, $\lambda_i \in [0.5, 1]$, kernel size=15, transformation function probability p=0.7. For the other baseline methods, we followed the best implementation from the original paper.

B. Eroding Gradients of Back Propagation

In the methodology section, we introduce our proposed erosion back-propagation gradient mechanism, proposing two approaches to the erosion of the back-propagation gradient of the residual network; see Fig. 1(c) and (d). Here, we will explore experimentally for both approaches, and we choose MXNET_LResNet50E_IR as our white-box model remaining other models as black-box models. The attack method uses the five commonly used adversarial attack methods mentioned before. The results are shown in Table II. The use of dynamic erosion form like Fig. 1(c) is not effective in enhancing transferability for various attack methods. On the contrary, the use of dynamic erosion form like Fig. 1(d) can substantially enhance the transferability for various attack methods. Up to 14.8% fooling rate can be improved. Therefore, we recommend using the dynamic erosion form like Fig. 1(d) for adversarial attacks.

C. Comparison With Ghost Network

Our dynamic erosion back-propagation gradient mechanism is compared with Ghost erosion, and the results are shown in Table I. Other adversarial attack methods are combined with our GE and Ghost, respectively. The MXNET_LResNet50E_IR model is used as a white-box model and other black-box models are attacked to verify the transferability. It can be seen that our method can substantially improve the transferability of the adversarial attack compared to Ghost. The reason lies in the fact that Ghost loses the accuracy of the model, resulting in a

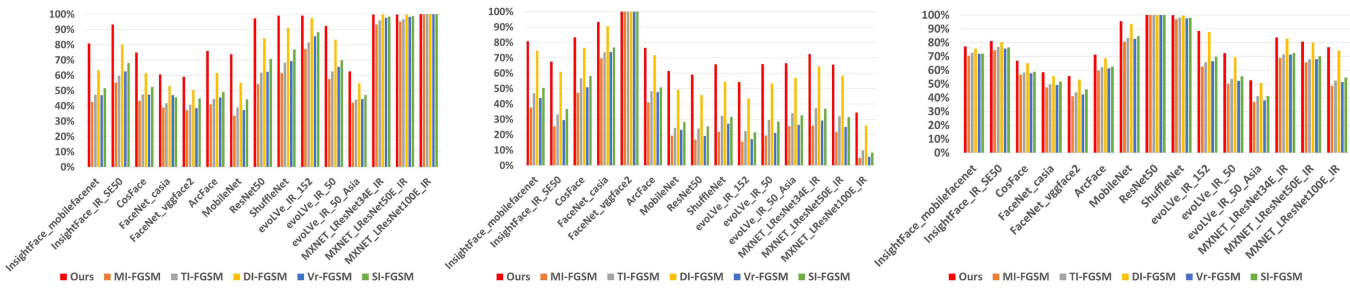
TABLE I
COMPARISON OF OUR DYNAMIC EROSION BACK-PROPAGATION GRADIENT MECHANISM WITH GHOST EROSION

Method	InsightFace_ mobilefacenet	InsightFace_ IR_SE50	CosFace	FaceNet_ casia	FaceNet_ vggface2	ArcFace	MobileNet	ResNet50	ShuffleNet	evoLVe_ IR_152	evoLVe_ IR_50	evoLVe_IR _50_Asia
MI-FGSM(Ghost)	**64.05%**	60.25%	56.35%	47.80%	33.55%	49.00%	34.25%	39.60%	56.75%	40.50%	36.3%	29.05%
MI-FGSM(Ours)	60.35%	**73.55%**	**60.35%**	**54.25%**	**52.00%**	**55.60%**	**46.60%**	**62.95%**	**72.60%**	**69.70%**	**61.25%**	**53.10%**
TI-FGSM(Ghost)	**67.95%**	63.5%	**62.60%**	52.25%	38.90%	54.10%	39.30%	47.65%	62.70%	44.35%	40.85%	32.55%
TI-FGSM(Ours)	63.80%	**75.80%**	62.35%	**55.55%**	**54.25%**	**57.95%**	**49.80%**	**66.65%**	**76.65%**	**71.70%**	**62.95%**	**54.35%**
DI-FGSM(Ghost)	66.75%	58.80%	55.95%	46.30%	34.55%	48.35%	34.95%	41.25%	58.40%	38.50%	33.65%	26.65%
DI-FGSM(Ours)	**92.05%**	**98.60%**	**85.15%**	**76.80%**	**74.20%**	**84.30%**	**75.05%**	**92.90%**	**98.15%**	**94.65%**	**86.60%**	**73.10%**
Vr-FGSM(Ghost)	**68.20%**	63.85%	59.80%	50.60%	36.95%	51.80%	37.45%	43.75%	61.35%	43.50%	39.50%	32.90%
Vr-FGSM(Ours)	67.25%	**82.00%**	**64.95%**	**58.25%**	**56.70%**	**61.45%**	**51.75%**	**70.30%**	**80.05%**	**77.75%**	**66.85%**	**56.85%**
SI-FGSM(Ghost)	**76.25%**	72.15%	**68.30%**	54.80%	42.05%	57.90%	45.40%	54.70%	71.05%	50.00%	45.20%	36.65%
SI-FGSM(Ours)	68.65%	**83.90%**	66.00%	**59.20%**	**58.85%**	**61.05%**	**54.55%**	**73.25%**	**82.8%**	**77.85%**	**68.10%**	**58.05%**

GE performs more consistently and gets a higher fooling rate compared to ghost.

TABLE II
EFFECT OF BACK-PROPAGATION GRADIENT EROSION ON THE FOOLING RATE OF FIVE ADVERSARIAL ATTACKS, WHERE (D) INDICATES THE GRADIENT EROSION MECHANISM SHOWN IN FIG. 1(D), (C) INDICATES THE GRADIENT EROSION MECHANISM SHOWN IN FIG. 1(C), AND NO ADDITION INDICATES THE ORIGINAL ATTACK METHOD

Method	InsightFace_ mobilefacenet	InsightFace_ IR_SE50	CosFace	FaceNet_ casia	FaceNet_ vggface2	ArcFace	MobileNet	ResNet50	ShuffleNet	evoLVe_ IR_152	evoLVe_ IR_50	evoLVe_IR _50_Asia
MI-FGSM(d)	**60.35%**	**73.55%**	**60.35%**	**54.25%**	**52.00%**	**55.60%**	**46.60%**	**62.95%**	**72.60%**	**69.70%**	**61.25%**	**53.10%**
MI-FGSM(c)	48.05%	63.55%	51.00%	45.25%	44.35%	47.65%	36.30%	54.10%	62.10%	66.50%	55.65%	45.15%
MI-FGSM	45.55%	62.90%	48.30%	43.25%	41.45%	45.50%	33.85%	51.40%	58.65%	65.50%	54.30%	42.65%
TI-FGSM(d)	**63.80%**	**75.80%**	**62.35%**	**55.55%**	**54.25%**	**57.95%**	**49.80%**	**66.65%**	**76.65%**	**71.70%**	**62.95%**	**54.35%**
TI-FGSM(c)	50.95%	66.95%	53.75%	46.55%	45.85%	50.05%	40.85%	59.05%	68.20%	69.30%	57.80%	46.60%
TI-FGSM	49.50%	66.40%	51.20%	45.20%	43.90%	47.60%	39.15%	57.10%	65.30%	68.20%	56.70%	44.95%
DI-FGSM(d)	**92.05%**	**98.60%**	**85.15%**	**76.80%**	**74.20%**	**84.30%**	**75.05%**	**92.90%**	**98.15%**	**94.65%**	**86.60%**	**73.10%**
DI-FGSM(c)	84.45%	96.25%	77.85%	67.15%	65.50%	78.40%	67.50%	90.05%	96.40%	93.80%	83.105%	67.30%
DI-FGSM	83.40%	96.45%	76.85%	64.70%	64.25%	78.20%	65.20%	89.70%	95.95%	93.90%	82.75%	65.15%
Vr-FGSM(d)	**67.25%**	**82.00%**	**64.95%**	**58.25%**	**56.70%**	**61.45%**	**51.75%**	**70.30%**	**80.05%**	**77.75%**	**66.85%**	**56.85%**
Vr-FGSM(c)	54.90%	73.95%	54.95%	48.55%	45.95%	53.50%	41.45%	62.95%	72.00%	75.45%	62.60%	49.15%
Vr-FGSM	52.55%	71.95%	52.25%	46.75%	43.95%	50.05%	38.75%	58.60%	68.25%	73.90%	60.40%	46.55%
SI-FGSM(d)	**68.65%**	**83.90%**	**66.00%**	**59.20%**	**58.85%**	**61.05%**	**54.55%**	**73.25%**	**82.80%**	**77.85%**	**68.10%**	**58.05%**
SI-FGSM(c)	59.85%	77.35%	58.60%	51.05%	50.80%	54.20%	46.20%	68.45%	76.70%	75.50%	64.15%	50.80%
SI-FGSM	56.35%	75.95%	55.75%	48.70%	48.80%	52.60%	42.95%	66.00%	73.15%	72.95%	61.75%	48.15%



(a) (b) (c)

Fig. 2. Comparison of corrasion attack with other five attack methods under different models, experiments set epsilon=20.4, iters=100, the number of subiteration rounds is 3, the horizontal axis denotes the models in our library of 15 face feature extraction models, and the vertical axis indicates the fooling rate. (a) Comparison of the effect under the white-box model MXNET_LRexNet100E_IR. (b) Comparison of the effect under the white-box model FaceNet_vggface2. (c) Comparison of the effect under the white-box model Resnet50. Our corrasion attack method performs higher transferability on unknown black-box models than other methods.

certain error between the model's prediction and ground truth. In contrast, our method only dynamically erodes the gradient during back-propagation, which does not affect the accuracy of the model.

D. Attacking Black-Box Models

Our method corrasion attack is compared here with five currently used attack methods (MI-FGSM, TI-FGSM, DI-FGSM, Vr-FGSM, and SI-FGSM). Since we aim at targeting attacks for FR tasks, we use cosine similarity as our loss function. The baseline of our attack method is set to the momentum-based MI-FGSM method. We use MXNET_LRESNet100E_IR, FaceNet_vggface2, and Resnet50 as white-box models, respectively. The rest of the models are used as unknown black-box models for our experiments, and the results are shown in Fig. 2. Our method obtained 83.45%, 67.6%, and 75.75% average black-box model fooling rates when using MXNET_LRESNet100E_IR, FaceNet_vggface2, and Resnet50 as white-box models, respectively, while the DI-FGSM method obtained 74%, 59%, and 74.05% average black-box model attack. The fooling rates other methods show are even lower. After comparing experiments in three single models, our approach has a higher fooling rate and transferability than the five counterattack methods commonly used today.

TABLE III
FOOLING RATE OF CORRASION ATTACK WITH FIVE ADVERSARIAL ATTACK METHODS ON THREE ENSEMBLE MODELS. OUR CORRASION ATTACK IS REPRESENTED BY "OURS" AND "(GE)" INDICATES THE COMBINATION OF A METHOD WITH AN GE EROSION MECHANISM, AND CORRASION ATTACK COMBINED WITH GE CAN IMPROVE THE FOOLING RATE OF MI-FGSM METHOD BY 29.35%

Method	InsightFace_ mobilefacenet	InsightFace_ IR_SE50	CosFace	FaceNet_ casia	FaceNet_ vggface2	ArcFace	MobileNet	ResNet50	ShuffleNet	evoLVe_ IR_152	evoLVe_ IR_50	evoLVe_IR _50_Asia
Ours	95.10%	99.20%	90.00%	79.95%	63.30%	90.70%	**83.95%**	**98.05%**	99.60%	**98.85%**	**94.65%**	**79.45%**
Ours(GE)	**96.15%**	**99.35%**	**90.75%**	**80.75%**	**79.05%**	**91.45%**	83.30%	97.75%	**99.65%**	97.90%	92.20%	78.50%
MI-FGSM	66.80%	84.45%	65.25%	58.10%	57.65%	63.15%	52.70%	77.30%	85.00%	90.25%	76.50%	59.60%
MI-FGSM(GE)	75.20%	90.80%	71.85%	63.45%	61.90%	68.20%	59.70%	82.75%	90.75%	91.80%	80.30%	64.65%
TI-FGSM	70.95%	87.40%	69.10%	60.00%	61.05%	66.90%	58.35%	82.25%	89.90%	92.90%	79.95%	62.25%
TI-FGSM(GE)	78.35%	92.90%	75.15%	65.55%	64.55%	71.35%	64.50%	87.25%	94.15%	93.15%	82.60%	66.85%
DI-FGSM	87.60%	97.65%	82.80%	73.30%	71.70%	83.70%	74.00%	93.65%	97.90%	98.25%	91.05%	73.70%
DI-FGSM(GE)	93.75%	99.00%	87.55%	77.00%	75.15%	87.85%	78.15%	96.05%	99.15%	97.95%	90.95%	75.95%
Vr-FGSM	70.50%	88.40%	68.50%	61.15%	60.10%	66.90%	57.45%	81.05%	88.25%	93.20%	80.80%	63.10%
Vr-FGSM(GE)	88.20%	97.85%	82.15%	72.60%	71.45%	81.15%	72.75%	93.60%	98.00%	97.35%	89.60%	73.50%
SI-FGSM	78.65%	93.40%	75.35%	64.75%	64.75%	71.70%	65.65%	88.80%	94.25%	95.05%	84.00%	66.35%
SI-FGSM(GE)	82.50%	95.85%	77.95%	67.90%	67.40%	73.95%	67.85%	90.25%	95.65%	94.95%	84.90%	69.00%

TABLE IV
ATTACK ON ONLINE FR SYSTEM

OnlineSystem	ACC	Objective
AWS	84.5%	Targeted
Clarifai	89.8%	Untargeted

Targeted attack on face comparison API for AWS online system. Nontargeted attack on the Clarifai online system for celebrity recognition.

TABLE V
DATA IN THE TABLE REPRESENTS THE PERCENTAGE OF ROBUSTNESS IMPROVEMENT OF THE MODEL FOR A ADVERSARIAL ATTACK METHOD

Model	MI-FGSM	TI-FGSM	DI-FGSM	Vr-FGSM	SI-FGSM
MobileFaceNet	58.60%	54.25%	26.50%	52.05%	43.90%
CosFace	45.90%	50.40%	19.35%	37.60%	33.60%

Adversarial training model is robust to all common adversarial attacks. With five common attack methods, we improve the robustness of the model by an average of 43.2%.

Ensemble models attacks are an important tool for improving the transferability of adversarial examples, and assembling multiple alternative models for adversarial attacks can substantially improve the fooling rate of attacking unknown black-box models. We use MXNET_LResNet34E_IR, MXNET_LResNet50E_IR, and MXNET_LResNet100 E_IR as white-box models for targeted attacks. We compare the black-box fooling rate results of our method with five adversarial attack methods in two different contexts of Fig. 1(d) and the base model as shown in Table III. It is easy to find that our method shows strong transferability in ensemble models attack compared to several other methods. Our method achieves an average fooling rate of 89.4% under 12 unknown black-box models, while the remaining five aggressive methods only have { MI-FGSM: 69.72%, TI-FGSM: 73.41%, DI-FGSM:85.44% Vr-FGSM:73.28%, SI-FGSM:78.55%}. When these methods were combined with our GE erosion mechanism, we took an average fooling rate of 82.90% for the six methods (five common adversarial methods with our corrasion attack). In contrast, in the unErode model, the average fooling rate of the six methods was 78.3%. Our GE mechanism can effectively improve the fooling rate of the black-box model attack for the adversarial attack methods.

Besides, we also perform black-box adversarial attacks on the APIs provided by two current online FR systems. The results are shown in Table IV. For the face comparison API in the AWS online system, we perform a targeted attack on it and we achieve an 84.5% fooling rate of black-box targeted attack. For the celebrity recognition API in the Clarifai online system, we perform an untargeted attack, and we achieve a black-box targeted fooling rate of 89.8%. This shows that our method can also guarantee a high fooling rate against the black-box model of the online recognition system, which in turn indicates that the adversarial samples generated by our method have high aggressiveness and high transferability.

E. Improving Model Robustness

We use our method to perform adversarial training [43] on two FR models, MobileFaceNet and CosFace. The adversarial training is performed on the CASIA-WebFace dataset containing 490 607 face images with 10 572 categories. The trained models are tested for model accuracy on LFW, the most widely used dataset for face verification benchmarks. The accuracy of the adversarially trained MobileFaceNet model on the LFW dataset is 98.63%, and the accuracy of the CosFace model on the LFW dataset is 97.79%. We use the five commonly used adversarial attack methods mentioned before to adversarial attack the trained MobileFaceNet and CosFace models, respectively, and the results are shown in Table V. The model improves the robustness of the model as well as the overall robustness of the model for different attack methods, which can improve the robustness of the model by up to 43.2%. This also indicates that our attack method learns more essential image features to make the model have stronger generalization ability.

V. DISCUSSION

Our research focuses on targeted attacks in FR tasks and on improving the transferability of adversarial examples. We believe that targeted attacks are more threatening and challenging in FR tasks. To obtain more diverse gradient information from alternative models to improve the transferability of adversarial examples, we propose two methods for dynamically eroding the gradients of residual networks. We show that dynamic erosion of gradients in back-propagation over residual blocks can significantly improve the transferability of adversarial samples. We consider several reasons for this.

1) We ensured the accuracy of the model predictions, rather than losing the model's accuracy as Ghost did, which led to bias in the loss calculation.
2) We obtain more gradient information from the lower layers by eroding the gradient information back-propagated from the residual blocks.

Our research focuses on networks with residual structures in FR tasks. We will concentrate on residual structured networks and other vision tasks in our future work.

VI. Conclusion

In this article, we propose GE to erode gradients in back-propagation dynamically. Our method can obtain more diverse gradients to improve the adversarial examples' transferability without affecting the model's accuracy. Combining GE with various adversarial attack methods can significantly enhance the fooling rate of black-box attacks. Proposed corrasion attack has higher transferability than other adversarial attack methods and can achieve a higher fooling rate when attacking unknown black-box models and online FR systems. Also, adversarial training of models using our method can improve the model's robustness by 43.2%.

Our research improves the transferability of adversarial examples further and improves the robustness of the model. In our future work, we will continue our research related to the robustness of the model.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 7-9, 2015.

[3] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[4] P. A. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2001.

[5] B. Wu, H. Ai, C. Huang, and S. Lao, "Fast rotation invariant multi-view face detection based on real adaboost," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2004, pp. 79–84, 2004.

[6] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1728–1740, Oct. 2007.

[7] T. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understanding*, vol. 61, pp. 38–59, 1995.

[8] X. Yang, F. Wei, H. Zhang, X. Ming, and J. Zhu, "Design and interpretation of universal adversarial patches in face detection," *CoRR*, vol. abs/1912.05021, 2019. [Online]. Available: http://arxiv.org/abs/1912.05021

[9] B. Yin et al., "Adv-makeup: A new imperceptible and transferable attack on face recognition," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Z. Zhou, Ed., ijcai.org, Virtual Event/Montreal, Canada, Aug. 19-27, 2021, pp. 1252–1258.

[10] Y. Wang, Y. Tan, T. Baker, N. Kumar, and Q. Zhang, "Deep fusion: Crafting transferable adversarial examples and improving robustness of industrial artificial intelligence of things," *IEEE Trans. Ind. Inform.*, to be published, doi: 10.1109/TII.2022.3168874.

[11] Y. Wang, Y. Tan, H. Lyu, S.-H. Wu, Y. Zhao, and Y. Li, "Toward feature space adversarial attack in the frequency domain," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 11 019–11 036, 2022. [Online]. Available: https://doi.org/10.1002/int.23031

[12] Y. Zhang, Y. Tan, M. Lu, T. Chen, Y. Li, and Q. Zhang, "Boosting cross-task adversarial attack with random blur," *Int. J. Intell. Syst.*, vol. 37, pp. 8139–8154, 2022.

[13] L. Haoran, T. Yu-an, X. Yuan, W. Yajie, and X. Jingfeng, "A CMA-ES-based adversarial attack against black-box object detectors," *Chin. J. Electron.*, vol. 30, pp. 406–412, 2021.

[14] Y. Wang et al., "Towards a physical-world adversarial patch for blinding object detection models," *Inf. Sci.*, vol. 556, pp. 459–471, 2021.

[15] Q. Zhang, W. Ma, Y. Wang, Y. Zhang, Z. Shi, and Y. Li, "Backdoor attacks on image classification models in deep neural networks," *Chin. J. Electron.*, vol. 31, no. 2, pp. 199–212, 2022.

[16] H. Sun, Y. Tan, L. Zhu, Q. Zhang, Y. Li, and S. Wu, "A fine-grained and traceable multidomain secure data-sharing model for intelligent terminals in edge-cloud collaboration scenarios," *Int. J. Intell. Syst.*, vol. 37, pp. 2543–2566, 2022.

[17] H. Sun, Y. Tan, C. Li, L. Lei, Q. Zhang, and J. Hu, "An edge-cloud collaborative cross-domain identity-based authentication protocol with privacy protection," *Chin. J. Electron.*, vol. 31, no. 4, pp. 721–731, 2022.

[18] J. Yang, J. Zheng, T. Baker, S. Tang, Y. Tan, and Q. Zhang, "Clean-label poisoning attacks on federated learning for IoT," *Expert Syst.*, p. e13161, 2022.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 7-9, 2015. [Online]. Available: http://arxiv.org/abs/1412.6

[20] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. 2nd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., Banff, AB, Canada, Apr. 14-16, 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

[21] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 86–94.

[22] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. 5th Int. Conf. Learn. Representations*, Workshop Track Proceedings, OpenReview.net, Toulon, France, Apr. 24-26, 2017. [Online]. Available: https://openreview.net/forum?id=HJGU3Rodl

[23] J. Su, D. V. Vargas, and S. Kouichi, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2017.

[24] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.

[25] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *CoRR*, vol. abs/1611.02770, 2016. [Online]. Available: http://arxiv.org/abs/1611.02770

[26] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9185–9193.

[27] C. Xie, Z. Zhang, J. Wang, Y. Zhou, Z. Ren, and A. Yuille, "Improving transferability of adversarial examples with input diversity," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2725–2734.

[28] L. Wu, Z. Zhu, C. Tai, and E. Weinan, "Understanding and enhancing the transferability of adversarial examples," *CoRR*, vol. abs/1605.07277, 2016. [Online]. Available: http://arxiv.org/abs/1605.07277

[29] N. Papernot, P. McDaniel, and I. J. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*.

[30] W. Zhou et al., "Transferable adversarial perturbations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 452–467.

[31] A. Kurakin et al., "Adversarial attacks and defences competition," *CoRR*, vol. abs/1804.00097, 2018. [Online]. Available: http://arxiv.org/abs/1804.00097

[32] Y. Li, S. Bai, Y. Zhou, C. Xie, Z. Zhang, and A. Yuille, "Learning transferable adversarial examples via ghost networks," in *Proc. Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11458–11465.

[33] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proc. 8th Int. Conf. Learn. Representations*, Addis Ababa, Ethiopia, OpenReview.net, Apr. 26-30, 2020. [Online]. Available: https://openreview.net/forum?id=SJlHwkBYDH

[34] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4307–4316.

[35] J. Deng, J. Guo, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4685–4694.

[36] H. Wang et al., "Cosface: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5265–5274.
[37] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.
[38] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. Biometric Recognit. 13th Chin. Conf.*, J. Zhao et al., Eds., Proceedings, ser. Lecture Notes in Computer Science, Urumqi, China, Springer, vol. 10996, Aug. 11-12, 2018, pp. 428–438. [Online]. Available: https://doi.org/10.1007/978-3-319-97909-0_46
[39] X. Yang, D. Yang, Y. Dong, W. Yu, H. Su, and J. Zhu, "Delving into the adversarial robustness on face recognition," *CoRR*, vol. abs/2007.04118, 2020. [Online]. Available: https://arxiv.org/abs/2007.04118
[40] Q. Wang, P. Zhang, H. Xiong, and J. Zhao, "Face.evolve: A high-performance face recognition library," *CoRR*, vol. abs/2107.08621, 2021. [Online]. Available: https://arxiv.org/abs/2107.08621
[41] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images: Detection, Alignment, Recognit.*, 2008.
[42] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning face representation from scratch," *CoRR*, vol. abs/1411.7923, 2014. [Online]. Available: http://arxiv.org/abs/1411.7923
[43] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. 6th Int. Conf. Learn. Representations*, Vancouver, BC, Canada, Conference Track Proceedings, OpenReview.net, Apr. 30–May 3, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

Huipeng Zhou is currently working toward the M.S. degree in artificial intelligence security and adversarial attacks against DNNs with the Beijing Institute of Technology, Beijing, China.

His research interests include machine learning and adversarial attack.

Yajie Wang is currently working toward the Ph.D. degree with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing, China.

His main research interests include the robustness and vulnerability of artificial intelligence, cyberspace security, etc.

Yu-an Tan received the B.S., M.S., and Ph.D. degrees in software and theory of computers from the Beijing Institute of Technology, Beijing, China, in 1991, 1994, and 2004, respectively.

Since 2010, he has been a Professor and Ph.D. supervisor with the Beijing Institute of Technology. His main research interests include information security, network storage, and embedded systems.

Dr. Tan is a Senior Member of the China Computer Federation.

Shangbo Wu received the B.S. degree from the Beijing Institute of Technology, in 2020, and the M.S. degree from the University of Glasgow, Glasgow, U.K., in 2022.

His main research interests include the areas of artificial intelligence security, specifically semantic black-box adversarial attacks.

Yuhang Zhao received the B.S. degree from the School of Electrical and Information, Northeast Agricultural University, Harbin, China, in 2019. He is currently working toward the Ph.D. degree majoring with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing, China .

His research interests include artificial intelligence security.

Quanxin Zhang received the Ph.D. degree in computer application technology from Beijing Institute of Technology, Beijing, China, in 2003.

He is currently an Associate Professor with the Beijing Institute of Technology. His current research interests include deep learning and information security.

Yuanzhang Li received the B.S., M.S., and Ph.D. degrees in software and theory of computer from the Beijing Institute of Technology, Beijing, China, in 2001, 2004, and 2015, respectively.

He is currently an Associate Professor with the Beijing Institute of Technology. His main research interests include mobile computing and information security.