Random Transformation of image brightness for adversarial attack

Bo Yang[a,b], Kaiyong Xu[a,b], Hengjun Wang[a,b] and Hengwei Zhang[a,b,*]
[a]*State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, Henan, China*
[b]*Henan Key Laboratory of Information Security, Zhengzhou, Henan, China*

Abstract. Deep neural networks (DNNs) are vulnerable to adversarial examples, which are crafted by adding small, human-imperceptible perturbations to the original images, but make the model output inaccurate predictions. Before DNNs are deployed, adversarial attacks can thus be an important method to evaluate and select robust models in safety-critical applications. However, under the challenging black-box setting, the attack success rate, i.e., the transferability of adversarial examples, still needs to be improved. Based on image augmentation methods, this paper found that random transformation of image brightness can eliminate overfitting in the generation of adversarial examples and improve their transferability. In light of this phenomenon, this paper proposes an adversarial example generation method, which can be integrated with Fast Gradient Sign Method (FGSM)-related methods to build a more robust gradient-based attack and to generate adversarial examples with better transferability. Extensive experiments on the ImageNet dataset have demonstrated the effectiveness of the aforementioned method. Whether on normally or adversarially trained networks, our method has a higher success rate for black-box attacks than other attack methods based on data augmentation. It is hoped that this method can help evaluate and improve the robustness of models.

Keywords: Adversarial examples, black-box attacks, deep neural networks (DNNs)

1. Introduction

In image recognition, some experiments on standard test sets have proven that deep neural networks (DNNs) have higher recognition ability than that of humans [1–4]. However, while deep learning brings great convenience, it also brings security problems. For an abnormal input, the question about whether DNNs can obtain satisfactory results remains. DNNs have been shown to be highly vulnerable to attacks from adversarial examples [5, 6], because adding perturbations to an original input image that are imperceptible to humans will cause misclassification of the models. As shown in Fig. 1, deep neural net-

works have been fooled into recognizing the Japanese spaniel as the great white shark. Furthermore, the experiments in [7] show that the "stop" sign added with human-imperceptible perturbations can deceive the neural network to identify it as a "speed limit 45" sign, which may mislead an autonomous vehicle to cause an accident. Specifically, adversarial examples normally have a certain degree of transferability, meaning those generated for one model may also be adversarial to another, which enables black-box attacks [8]. These phenomena show that the existence of transferable adversarial examples poses a great threat to the security of AI systems, leading to the chaos of AI driven intelligent systems, the formation of missed judgments and misjudgments, and even the collapse of the system. Therefore, it is particularly significant and urgent to study the reason for and essence of the existence of adversarial

*Corresponding author. Hengwei Zhang, State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, Henan, 450001, China. E-mail: wlby_zzmy_henan@163.com.
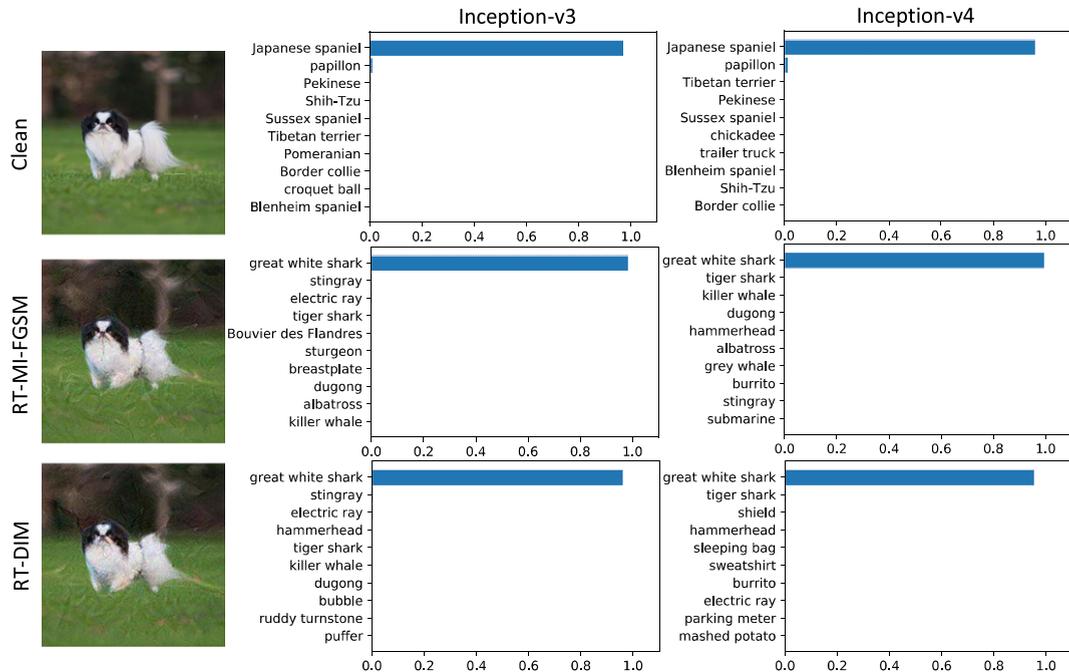
Fig. 1. The classification of a clean image and corresponding adversarial examples on Inception v3 and Inception v4 model are shown. For the images, the ground-truth is Japanese spaniel. The first row shows the top-10 confidence distributions for the clean image, which indicates all the models provide right prediction with high confidences. The second and third rows show the top-10 confidence distributions of the adversarial examples generated on the ensemble of models by RT-MI-FGSM and RT-DIM, which attack the two models successfully.

examples, as well as adversarial attack and adversarial defense. As to adversarial attack, it can be used to evaluate and test the robustness of deep neural networks; moreover, the adversarial examples generated by adversarial attack can be added to the training set for adversarial training, so as to enhance the robustness of models. Therefore, this paper is committed to the research of adversarial attack methods to help evaluate and improve the robustness of models.

Although adversarial examples are generally transferable, to further improve their transferability for effective black-box attacks remains to be explored. In the search of more transferable adversarial examples, some gradient-based attacks have been proposed, such as single-step [6] and iterative [9, 10] methods. These methods show powerful attack capabilities in the white-box setting, but their success rates are relatively low in the black-box setting, which is attributed to overfitting of adversarial examples. Since the generation process of adversarial examples is similar to the training process of neural network, this difference in attack ability of an adversarial example under white-box and black-box settings is also similar to that of the same neural network on training and test sets. As a result, this paper can apply methods that improve the performance of deep learning models to

the generation of adversarial examples to eliminate overfitting and improve their transferability. Many methods have been proposed to improve DNN performance [1, 2, 10–13], one of the most important one is data augmentation [1, 2], and it can prevent overfitting during training and improve the generalization ability of models.

This paper optimizes the generation of adversarial examples based on data augmentation and proposes the Random Transformation of Image Brightness Attack Method (RTM) to improve their transferability.

– Inspired by data augmentation [1, 2], this paper adapts the random transformation of image brightness to adversarial attacks, so as to effectively eliminate overfitting in the generation of adversarial examples and improve their transferability.

– The proposed method is readily combined with gradient-based attack methods (e.g., momentum iterative gradient-based [10] and diverse input [15] methods) to further boost the success rate of adversarial examples for black-box attacks.

Extensive experiments on the ImageNet dataset [14] have indicated that, compared to current data

augmentation attack methods [15], our method, RT-MI-FGSM (Random Transformation of Image Brightness Momentum Iterative Fast Gradient Sign Method), has a higher success rate for black-box attacks in normally and adversarially trained models. By Integrating RT-MI-FGSM with the diverse input method (DIM) [15], the resulting RT-DIM (Random Transformation of image brightness with Diverse Input Method) can greatly improve the average attack success rate on adversarially trained models in black-box settings. In addition, the method of attacking ensemble models simultaneously is used to further improve the transferability of adversarial examples [8]. Under the ensemble attack, RT-DIM reaches an average success rate of 72.1% for black-box attacks on adversarially trained networks, which outperforms DIM by a large margin of 24.3%. It is expected that the proposed attack method can help evaluate the robustness of models and effectiveness of defense methods.

2. Related work

2.1. Adversarial example generation

Biggio et al. [16] presented a simple but effective gradient-based method that can be used to systematically assess the security of several widely-used classification algorithms against evasion attacks, indicating that traditional machine learning models are vulnerable to adversarial examples. However, this discovery is limited to the traditional machine learning models, and cannot be extended to the widely used deep neural networks. Szegedy et al. [5] reported the intriguing property that DNNs are also fragile to adversarial examples and proposed the L-BFGS method to generate them, but this method needed a lot of computation. Goodfellow et al. [6] demonstrated the fast gradient sign method that can generate adversarial examples with one gradient step which reduces the computation needed to generate adversarial examples and forms the basis of subsequent FGSM-related methods, but has low attack success rate. Alexey et al. [9] extended FGSM to an iterative version, which greatly improved the success rate for white-box attacks and proved that adversarial examples also exist in the physical world. However, due to overfitting, the success rate of black box attack of this method is lower than that of FGSM. Dong et al. [10] proposed momentum-based iterative FGSM, improving the transferability of adversarial

examples. But this method only introduced a better optimization algorithm to generate adversarial examples, which limited the transferability of the adversarial examples. Zhang et al. [17] proposed a new approach named PCD for computing adversarial examples for DNNs and increase the robustness of Big Data. Because of the particularity of this method, this method cannot be well combined with FGSM-related methods and therefore, cannot further improve the attack success rate. Xie et al. [15] randomly transformed the original input images in each iteration to reduce overfitting and improved the transferability of adversarial examples. However, the realization of this method was not easy since random transformation involved scaling and adding. Dong et al. [18] used a set of translated images to optimize adversarial perturbations. To reduce computation, the gradient was calculated by convolving the gradient of the untranslated images with the kernel matrix, which can generate adversarial examples with better transferability. However, this method greatly increased the number of translation transformation, resulting in its significantly lower attack success rate on the normal training network than DIM. After discussing the above-mentioned methods, the fact that adversarial examples may exist in the physical world brings much greater security threats to the practical application of DNNs [7, 9].

2.2. Defense methods against adversarial examples

Many defense methods against adversarial examples have been proposed to protect deep learning models [19–26]. Adversarial training [6, 27, 28] is one of the most effective ways to improve the robustness of models by injecting adversarial examples into training data. Xie et al. [21] found that the effectiveness of adversarial examples can be reduced through random transformation. Guo et al. [22] found a range of image transformations with the potential to remove adversarial perturbations while preserving the key visual information of an image. Samangouei et al. [23] used a generative model to purify adversarial examples by moving them back toward the distribution of the original clean image, thereby reducing their impact. Liu et al. [24] proposed a JPEG-based defensive compression framework that can rectify adversarial examples without affecting classification accuracy on benign data, alleviating the adversarial effect. Cohen et al. [26] proposed a randomized smoothing technique to obtain an ImageNet classi-

fier with certified adversarial robustness. Tramèr et al. [28] proposed ensemble adversarial training, utilizing adversarial examples generated for other models to increase training data and further improve the robustness of models. Liu et al. [29] proposed a novel defense network based on generative adversarial network (GAN) to improve the robustness of the neural network.

3. Methodology

Let x be the original input image, y the corresponding true label, and θ the parameter of the model. $J(\theta, x, y)$ is the loss function of the neural network, which is usually cross-entropy loss. This paper aims to generate an adversarial example x^{adv} that is visually indistinguishable from x by maximizing $J(\theta, x, y)$ to fool the model; i.e., the model misclassifies the adversarial example x^{adv}. This paper uses the L_{∞} norm bound, $||x^{adv} - x||_{\infty} \leq \varepsilon$, to limit adversarial perturbations. Hence adversarial example generation can be transformed to a condition constrained optimization problem:

$$\arg\max_{x^{adv}} J(\theta, x^{adv}, y), \quad s.t. ||x^{adv} - x||_{\infty} \leq \varepsilon. \quad (1)$$

3.1. Gradient-based adversarial attack methods

Since our proposed method belongs to and is based on gradient-based adversarial attack methods, this section introduces several methods to generate adversarial examples.

Fast Gradient Sign Method (FGSM). FGSM [6] is one of the most basic methods, which searches adversarial examples in the direction of the loss gradient $\nabla_x J(\theta, x, y)$ with respect to the input and imposes infinity norm restrictions on adversarial perturbations. The updated equation is

$$x^{adv} = x + \varepsilon \cdot sign(\nabla_x J(\theta, x, y)). \quad (2)$$

Iterative Fast Gradient Sign Method (I-FGSM). Kurakin et al. [9] proposed an iterative version of FGSM. It divides the gradient operation in FGSM into multiple iterations to eliminate the under-fitting caused by single-step attacks. It can be expressed as

$$x_{t+1}^{adv} = \text{Clip}_x^{\varepsilon}\{x_t^{adv} + \alpha \cdot sign(\nabla_x J(\theta, x_t^{adv}, y))\}, \quad (3)$$

where α is the step size of each iteration and $\alpha = \varepsilon/T$, where T is the number of iterations. The Clip function

restricts the adversarial example to be within the ε-ball of the original image x to meet the infinity norm constraint. Experiments have shown that I-FGSM has a higher success rate for white-box attacks than FGSM, but with poorer transferability.

Momentum Iterative Fast Gradient Sign Method (MI-FGSM). Dong et al. [10] proposed MI-FGSM from the perspective of introducing better optimization algorithm to the process of adversarial attack. MI-FGSM is the first method to integrate the momentum item with I-FGSM, which can stabilize gradient update directions, improve convergence, and greatly increase the attack success rate. MI-FGSM differs from I-FGSM in the update directions of adversarial examples:

$$x_0^{adv} = x, \ g_0 = 0, \ g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(\theta, x_t^{adv}, y)}{||\nabla_x J(\theta, x_t^{adv}, y)||_1}, \quad (4)$$

$$x_{t+1}^{adv} = \text{Clip}_x^{\varepsilon}\{x_t^{adv} + \alpha \cdot sign(g_{t+1})\}, \quad (5)$$

where g_t is the gradient weighted accumulation of the previous t iterations. In Equation (4), μ is the decay factor of the momentum term, which controls momentum updating size.

Diverse Input Method (DIM). DIM [15] randomly transforms the original input with a given probability in each iteration to reduce overfitting. Transformations include random resizing and padding. This method is readily combined with other baseline attack methods to generate adversarial examples with better transferability. The random transformation equation is

$$T(X_t^{adv}; p) = \begin{cases} T(X_t^{adv}), \text{with probability } p \\ X_t^{adv}, \text{with probability } 1 - p \end{cases}. \quad (6)$$

Projected Gradient Descent (PGD). PGD [19] is a strong iterative version of FGSM, which improves the attack success rate of adversarial examples.

3.2. Random transformation of image brightness attack method

Data augmentation [1, 2] has been proven effective to prevent network overfitting during DNN training. Based on this, this paper proposes the Random Transformation of Image Brightness Attack Method (RTM), which randomly transforms the brightness of the original input image with probability p in each iteration to alleviate overfitting. The idea of the RTM
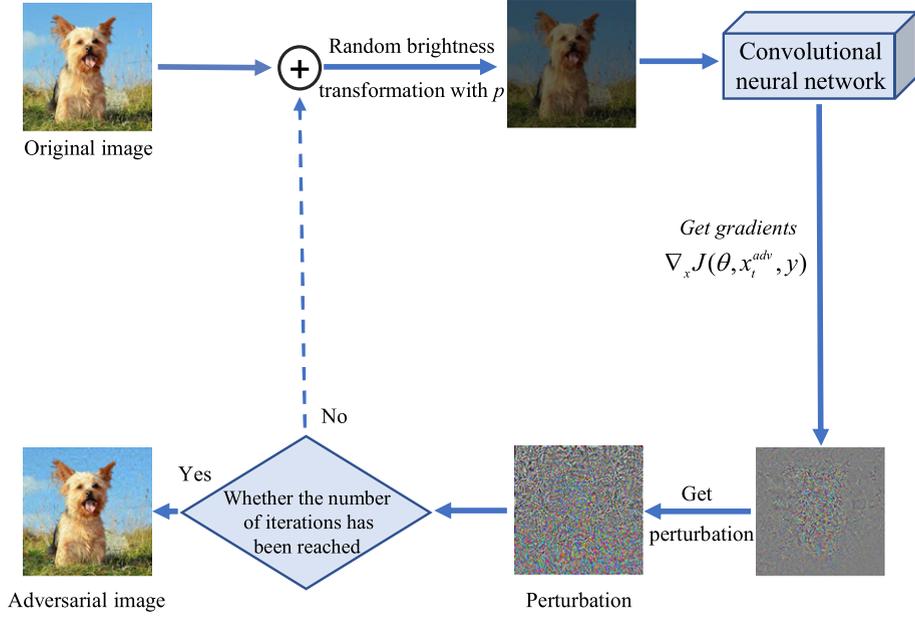
Fig. 2. Frame diagram of random transformation of image brightness attack method.

is showed on Fig. 2. It optimizes the adversarial perturbations of the image with randomly transformed brightness:

$$\arg\max_{x^{adv}} J(\theta, RT(x^{adv}; p), y), \ \text{s.t.} ||x^{adv} - x||_\infty \leq \varepsilon, \quad (7)$$

$$RT(X^{adv}; p) = \begin{cases} RT(X^{adv}) m \ \text{with} \ \text{probability} \ p \\ X^{adv}, \text{with} \ \text{probability} \ 1 - p \end{cases}. \quad (8)$$

The random transformation function $RT(\cdot)$ randomly decreases the brightness of the input image at a random rate. The transformation probability p controls the balance between the original input image and the transformed image. With this method, this paper can achieve effective attacks on the model through data augmentation, avoid overfitting attacks of white-box models, and improve the transferability of adversarial examples.

3.3. Attack algorithms

For the gradient processing of generating adversarial examples, RTM introduces data augmentation to alleviate overfitting. RTM is easily combined with MI-FGSM to form a stronger attack, which is referred to as RT-MI-FGSM (Random Transformation of image brightness Momentum Iterative Fast Gradient Sign Method). Our algorithm can be associated with the family of FGSM by adjusting its param-

---

Algorithm 1: RT-MI-FGSM

Input: A clean example x with ground-truth label y; a classifier f with loss function J; Perturbation size ε; maximum iterations T and decay factor μ.

Output: An adversarial example x^{adv}

1: $\alpha = \varepsilon / T$
2: $x_0^{adv} = x$; $g_0 = 0$
3: **for** $t = 0$ to $T - 1$ **do**
4:    Get x_t^{adv} by $x_t^{adv} = RT(x_t^{adv}; p) \triangleright$ apply random transformation of the input's brightness with the probability p
5:    Get the gradients by $\nabla_x J(\theta, x_t^{adv}, y)$
6:    Update g_{t+1} by $g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(\theta, x_t^{adv}, y)}{||\nabla_x J(\theta, x_t^{adv}, y)||_1}$
7:    Update x_{t+1}^{adv} by Equation (5)
8: **return** $x^{adv} = x_T^{adv}$

---

eter settings. For example, RT-MI-FGSM degrades to MI-FGSM if $p = 0$, i.e., step 4 of algorithm 1 can be removed to realize MI-FGSM. Algorithm 1 summarizes the RT-MI-FGSM attack algorithm[1].

In addition, RTM can be combined with DIM to form RT-DIM, further improving the transferability of adversarial examples. The algorithm of RT-DIM attack is summarized in Algorithm 2. The RT-MI-FGSM attack algorithm can be obtained by removing step 4 of Algorithm 2 and the DIM attack algorithm by removing step 5. In addition, the MI-FGSM attack

---

[1]The code is publicly available at https://github.com/yangbo93/atack-method.

Algorithm 2: RT-DIM

Input: A clean example x with ground-truth label y; a classifier f with loss function J; Perturbation size ε; maximum iterations T and decay factor μ.
Output: An adversarial example x^{adv}

1: $\alpha = \varepsilon / T$
2: $x_0^{adv} = x; g_0 = 0$
3: **for** $t = 0$ to $T - 1$ **do**
4:    Get x_t^{adv} by $x_t^{adv} = RT(x_t^{adv}; p) \triangleright$ apply random transformation of the input's brightness with the probability p
5:    Get x_t^{adv} by $x_t^{adv} = T(x_t^{adv}; p) \triangleright$ apply random resizing and padding to the inputs with the probability p
6:    Get the gradients by $\nabla_x J(\theta, x_t^{adv}, y)$
7:    Update g_{t+1} by $g_{t+1} = \mu \cdot g_t + \frac{g}{\|g\|_1}$
8:    Update x_{t+1}^{adv} by Equation (5)
9: **return** $x^{adv} = x_T^{adv}$

---

algorithm can be acquired by removing steps 4 and 5. Of course, our method can also be related to the family of Fast Gradient Sign Methods by adjusting the transformation probability p and decay rate μ. It reflects the convenience and advantages of our method.

4. Experiments

Extensive experiments are conducted to evaluate our method's effectiveness. In the following, we specify the experimental settings, show the results of attacking a single network, validate our method on ensemble models, and discuss the hyper-parameters that affect the results.

4.1. Experimental setup

Dataset. It is meaningless to generate adversarial examples from the original images that are already classified wrongly. This paper randomly selects 1000 images belonging to 1000 categories (i.e., one image per category) from the ImageNet verification set, which were correctly classified by our testing networks. All images were adjusted to $299 \times 299 \times 3$.

Models. Seven networks are considered. The four normally trained networks are Inception-v3 (Inc-v3) [30], Inception-v4 (Inc-v4) [31], Inception-Resnet-v2 (IncRes-v2) [31], and Resnet-v2-101 (Res-101) [32]; the three adversarially trained networks [28] are ens3-adv-Inception-v3 (Inc-v3$_{ens3}$), ens4-adv-Inception-v3 (Inc-v3$_{ens4}$), and ens-adv-Inception-ResNet-v2 (IncRes-v2$_{ens}$).

Baselines. Our method is integrated with MI-FGSM [10] and DIM [15] to evaluate the improvement of RTM over these baseline methods.

Implementation details. For the hyper-parameters, we follow the default settings in [15] with the maximum perturbation $\varepsilon = 16$, number of iterations $T = 10$, and step size $\alpha = 1.6$. For MI-FGSM, the decay factor is defaulted to $\mu = 1.0$. For DIM, the default settings are adopted. For our methods, p is set to 0.5 for the random transformation function $RT(X; p)$, and to 1.0 when RTM is combined with DIM. For transformation operations $RT(\cdot)$, the brightness of the input image x is randomly adjusted to $B * r$, where B is the original brightness of the input image and $r \in (1/16, 1]$ is the adjustment rate. For intuitive understanding, Fig. 3 shows some images after random brightness transformation.

4.2. Attacking a single network

We first perform adversarial attacks on a single network. I-FGSM, MI-FGSM, DIM, and RT-MI-FGSM are used to generate adversarial examples only on the normally trained networks which are tested on all seven networks. The results are shown in Table 1, where the success rate is the model classification error rate with adversarial examples as input. This paper also combines RTM and DIM as RT-DIM. p is set to 1.0 for RTM in this case. The test results on the seven networks are shown in Table 2.

The results in Table 1 show that the attack success rates of RT-MI-FGSM under mostly black-box settings are much higher than those of other baseline attacks. It also has higher attack success rates than the DIM attack method based on data augmentation, and maintains relatively high white-box attack success rates. For example, when generating adversarial examples on the Inc-v3 network to attack the Inc-v4 network, the success rate for black-box attacks of RT-MI-FGSM reaches 71.4%, the highest among these methods. RT-MI-FGSM also performs better on the adversarially trained networks. Compared to the other three attack methods, our method greatly improves the success rates for black-box attacks. For example, when generating adversarial examples on the Inc-v3 network to attack the adversarially trained networks, the average attack success rates of RT-MI-FGSM and MI-FGSM are 24.6% and 12.2%, respectively. This 12.4% enhancement demonstrates that our method can effectively improve the transferability of adversarial examples. The six randomly selected original images, the corresponding randomly transformed images and generated adversarial examples are shown in Fig. 3. The adversarial examples are crafted on the Inc-v3 by the proposed RT-MI-FGSM

Fig. 3. The adversarial examples are crafted on Inc-v3 by RT-MI-FGSM and RT-DIM method respectively. Images from first to third line are original inputs, randomly transformed images, and generated adversarial examples, respectively.

Table 1
The success rates (%) of adversarial attacks against seven models under single model setting

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	I-FGSM	**99.9***	22.6	20.2	18.1	7.2	7.6	4.1
	MI-FGSM	**99.9***	48.1	47.1	39.9	15.2	14.2	7.2
	DIM	99.2*	69.6	64.8	58.8	22.7	21.2	10.3
	RT-MI-FGSM(**Ours**)	96.8*	**71.4**	**68.1**	**62.9**	**30.8**	**28.3**	**14.6**
Inc-v4	I-FGSM	37.9	**99.9***	26.2	21.9	8.7	8.0	5.0
	MI-FGSM	63.9	**99.9***	53.7	47.7	19.7	16.9	9.4
	DIM	80.1	99.0*	71.4	63.6	26.6	24.9	13.4
	RT-MI-FGSM(**Ours**)	**80.4**	98.5*	**72.5**	**69.0**	**42.6**	**39.1**	**23.4**
IncRes-v2	I-FGSM	37.2	31.8	**99.6***	25.9	8.9	7.5	4.9
	MI-FGSM	68.6	61.9	**99.6***	52.1	25.1	20.2	14.4
	DIM	80.6	**76.5**	98.0*	69.7	36.6	32.4	22.6
	RT-MI-FGSM(**Ours**)	80.9	75.3	96.4*	**71.3**	**48.0**	**41.6**	**33.4**
Res-101	I-FGSM	27.7	23.3	21.3	**98.2***	9.3	7.9	5.6
	MI-FGSM	52.4	48.2	45.6	**98.2***	22.3	18.6	11.8
	DIM	**71.0**	**65.1**	**62.6**	97.5*	32.4	29.8	17.9
	RT-MI-FGSM(**Ours**)	66.5	61.8	59.7	96.7*	**33.7**	**29.9**	**20.3**

Adversarial examples are crafted on Inc-v3, Inc-v4, IncRes-v2, and Res-101, respectively, using I-FGSM, MI-FGSM, DIM, and RT-MI-FGSM. * indicates white-box attacks.

and RT-DIM method respectively. It can be seen that these generated adversarial perturbations are human imperceptible.

We then compare the attack success rates of the RT-MI-FGSM with that of DIM methods based on data augmentation. The results show that our method mostly performs better on both normally and adversarially trained networks, and RT-MI-FGSM has higher black-box attack success rates than DIM. In particular, compared to DIM, RT-MI-FGSM significantly improves the black-box attack success rates on the adversarially trained networks. For exam-

Table 2
The success rates (%) of adversarial attacks against seven models under single-model setting

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	DIM	**99.2***	69.6	64.8	58.8	22.7	21.2	10.3
	RT-DIM(**Ours**)	93.9*	**75.3**	**72.0**	**70.4**	**38.9**	**35.5**	**19.0**
Inc-v4	DIM	80.1	**99.0***	71.4	63.6	26.6	24.9	13.4
	RT-DIM(**Ours**)	**85.7**	96.1*	**77.5**	**74.9**	**54.0**	**48.3**	**34.1**
IncRes-v2	DIM	**80.6**	**76.5**	**98.0***	69.7	36.6	32.4	22.6
	RT-DIM(**Ours**)	77.6	73.4	89.6*	**70.0**	**51.1**	**45.8**	**38.0**
Res-101	DIM	**71.0**	65.1	62.6	**97.5***	32.4	29.8	17.9
	RT-DIM(**Ours**)	70.8	**65.5**	**62.9**	94.0*	**40.2**	**36.8**	**25.2**

Adversarial examples are crafted on Inc-v3, Inc-v4, IncRes-v2, and Res-101, respectively, using DIM and RT-DIM. * indicates white-box attacks.

ple, when generating adversarial examples on the Inc-v3 network to attack the adversarially trained network Inc-v3$_{ens4}$, the black-box attack success rate of DIM was 21.2%, and that of RT-MI-FGSM was 28.3%. If adversarial examples are crafted on Inc-v4, then RT-MI-FGSM has success rates of 42.6% on Inc-v3$_{ens3}$, 39.1% on Inc-v3$_{ens4}$, and 23.4% on IncRes-v2$_{ens}$, while DIM only obtains corresponding success rates of 26.6%, 24.9%, and 13.4%, respectively.

The results in Table 2 show that RT-DIM, which integrates RT-MI-FGSM and DIM, further improves the attack success rates in most black-box settings. For example, when generating adversarial examples on the Inc-v4 network to attack adversarially trained networks, the average attack success rate of RT-DIM reaches 45.5%, while that of the DIM method under the same conditions is 21.6%. The average attack success rate more than doubled with RT-DIM. Interestingly, the white-box attack success rates of RT-DIM are not as high as those of DIM, perhaps because the integration of the two methods further increases the transformation randomness of the original input image. More analysis and discussion about this can be seen in Section 4.4.

4.3. Attacking an ensemble of networks

Though RT-MI-FGSM and RT-DIM can improve the transferability of adversarial examples on the black-box models, we can further increase their attack success rates by attacking the ensemble models. We follow the strategy in [10] to attack multiple networks simultaneously. We consider all seven networks discussed above. Adversarial examples are crafted on an ensemble of six networks, and are tested on the ensembled network and hold-out network, using I-FGSM, MI-FGSM, DIM, RT-MI-FGSM, and RT-DIM, respectively. The number of iterations in the

iterative method is $T = 10$, the perturbation size is $\varepsilon = 16$, and the ensemble weights of networks are equal, i.e., $\omega_k = 1/6$.

The experimental results are summarized in Table 3, which shows that in the black-box settings, RT-DIM has higher attack success rates than the other methods. For example, with Inc-v3 as a hold-out network, the success rate of RT-DIM attacking Inc-v3 is 85.2%, while those of I-FGSM, MI-FGSM, DIM, and RT-MI-FGSM are 54.3%, 75.4%, 83.7%, and 84.3%, respectively. On challenging adversarially trained networks, the average success rate of RT-DIM for black-box attacks is 72.1%, which is 24.3% higher than that of DIM. These results show the effectiveness and advantages of our method.

In the white-box settings, we encounter a similar result to that of RT-MI-FGSM and RT-DIM mentioned above (see Section 4.2): the white-box attack success rates of RT-MI-FGSM on the ensemble models are lower than those of MI-FGSM, but are higher than those of RT-DIM and DIM, and the results of RT-DIM are lower than those of DIM and RT-MI-FGSM. This is an interesting result. Perhaps the model and method ensembles have something in common, meaning they have similar effects on the generation of adversarial examples, which remains an open issue for future research.

4.4. Hyper-parameter studies

In this section, extended experiments are conducted to further study the influence of different parameters on RT-MI-FGSM and RT-DIM. This paper considers attacking an ensemble of networks to evaluate the robustness of the models more accurately [15]. The experimental settings are maximum perturbation $\varepsilon = 16$, number of iterations $T = 10$, and step size $\alpha = 1.6$. For MI-FGSM, the decay factor is

Table 3
The success rates (%) of adversarial attacks against seven models under multi-model setting

	Attack	-Inc-v3	-Inc-v4	-IncRes-v2	-Res-101	-Inc-v3$_{ens3}$	-Inc-v3$_{ens4}$	-IncRes-v2$_{ens}$
Ensemble	I-FGSM	98.5	98.2	99.1	**97.8**	98.2	98.2	96.3
	MI-FGSM	**98.8**	**98.7**	**99.4**	**97.8**	**98.6**	**98.5**	**96.8**
	RT-MI-FGSM(**Ours**)	93.9	95.4	96.2	93.5	95.5	95.0	96.3
	DIM	87.2	86.8	87.7	87.8	91.9	91.4	91.8
	RT-DIM(**Ours**)	87.7	87.8	90.0	89.0	91.3	92.8	91.9
Hold-out	I-FGSM	54.3	48.3	48.8	41.2	17.4	18.5	10.8
	MI-FGSM	75.4	69.7	67.5	62.8	25.4	31.2	19.1
	RT-MI-FGSM(**Ours**)	84.3	80.9	80.7	77.1	54.0	57.8	40.7
	DIM	83.7	82.3	80.5	76.9	49.8	52.0	41.7
	RT-DIM(**Ours**)	**85.2**	**82.4**	**84.0**	**81.9**	**75.2**	**75.7**	**65.4**

The "-" symbol indicates the name of the hold-out network. Adversarial examples are generated on the ensemble of the other six networks. The first row shows success rates for the ensembled networks (white-box attack), and the second row shows success rates for the hold-out network (black-box attack).
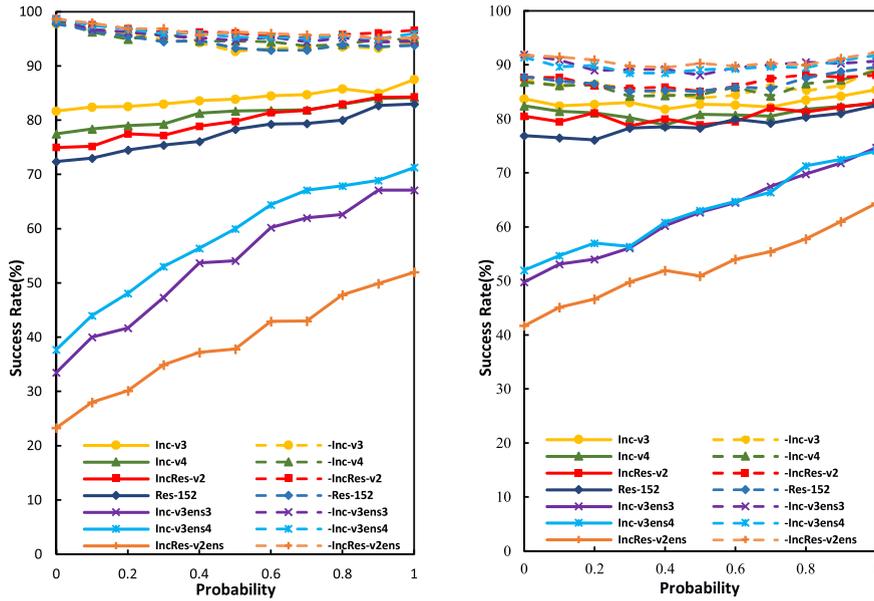

Fig. 4. Success rates of RT-MI-FGSM (left) and RT-DIM (right) under different transformation probabilities p. Adversarial examples are generated on an ensemble of six networks, and tested on the ensembled network (white-box setting, dashed line) and hold-out network (black-box setting, solid line).

defaulted to $\mu = 1.0$, and for DIM [15], the default settings are used.

Transformation probability p. This paper first studies the impact on the attack success rates in the white-box and black-box models by varying p between 0 and 1. When $p = 0$, RT-MI-FGSM degrades to MI-FGSM, and RT-DIM to DIM. Figure 4 shows the attack success rates of our method on various networks. It can be seen that the trends of RT-MI-FGSM and RT-DIM are different with the increase of p. For RT-MI-FGSM, as p increases, the success rates for black-box attacks increase and that for white-box attacks decrease. For RT-DIM, with the increase of p, the success rates of black

box attack on adversarially trained networks increase steadily. The success rates of white-box attacks and black-box attacks on normally trained network first increase and then decrease, and finally present an upward trend. Moreover, for all attacks, if p is relatively small, i.e., only utilizing a small number of randomly transformed inputs, the black-box success rates on adversarially trained models increase significantly, the black-box success rates on normally trained networks fluctuate slightly, while the white-box success rates only drop a little. This result shows the importance of adding randomly transformed inputs to the generation of adversarial examples.
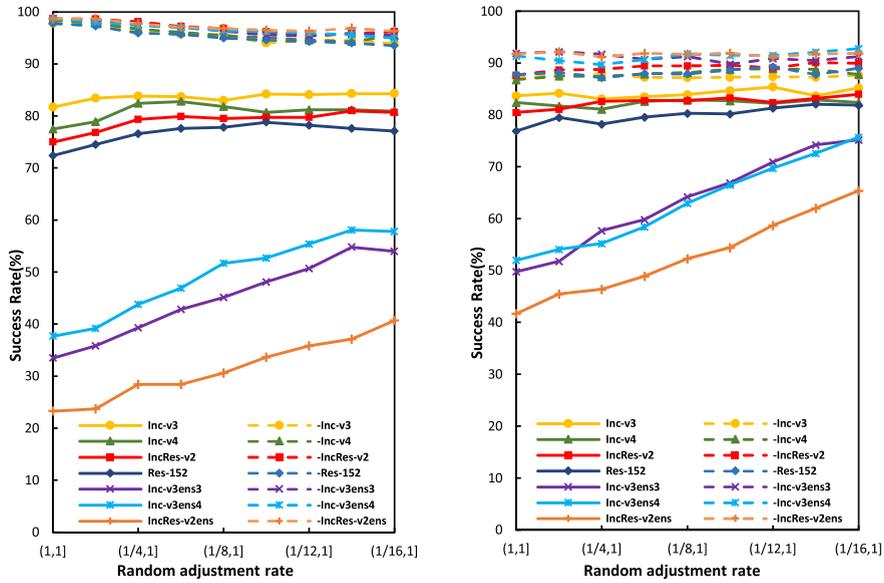
Fig. 5. Success rates of RT-MI-FGSM (left) and RT-DIM (right) under different random adjustment rates *r*. Adversarial examples are generated on an ensemble of six networks and tested on the ensembled network (white-box setting, dashed line) and hold-out network (black-box setting, solid line).

Random adjustment rate r. This paper studies the effect of *r* on the attack success rates under the white-box and black-box settings, randomly selecting *r* within a range, which is changed from (1, 1] to $\left(1/16, 1\right]$. When $r \in (1, 1]$, i.e., $r = 1$, RT-MI-FGSM degrades to MI-FGSM, and RT-DIM degrades to DIM. The attack success rates on various networks are shown in Fig. 5. It can be seen that as the value range of *r* increases, the success rates for black-box attacks of RT-MI-FGSM increase, and they decrease for white-box attacks. However, for RT-DIM, the success rates of black-box attacks on adversarially trained networks are significantly improved, and the success rates of white-box attack and black-box attacks on normally trained networks are slightly increased. The random adjustment rate *r* considers both the randomness and adjustment amplitudes of the image brightness transformation.

Constant adjustment rate r. This paper finally studies the influence of a constant value of *r* on the attack success rates under the white-box and black-box settings, changing the value range of *r* from $1/16$ to 1, i.e., the value of *r* is increasing. In each iteration, the transformation of the original input is the same. When $r = 1$, RT-MI-FGSM degrades to MI-FGSM, and RT-DIM to DIM. Figure 6 shows the attack success rates on various networks. It should be noted

that the leftmost ordinate value in Fig. 6 represents the attack success rate of our method (RT-MI-FGSM and RT-DIM), while the ordinate value at the far right of Figs. 4 and 5 represents the attack success rate of our method. As *r* increases, the attack success rates of RT-MI-FGSM and RT-DIM have different trends. For RT-MI-FGSM, with the increase of *r*, i.e., the decrease of the amplitude of image brightness transformation, the success rates of black box attacks on adversarially trained networks decrease significantly. The success rates of black box attacks on normally trained networks first increase and then decrease, and finally present a downward trend, while the white-box attack success rates increase slightly. For RT-DIM, the success rates for black-box attacks on the adversarially trained networks decrease greatly, while they drop a little for white-box and black-box attacks on the normally trained networks. It is also found that the same methods have different attack effects under random and constant adjustment rates. With a random adjustment rate, a higher attack success rate is more easily achieved in the white-box models. With a constant adjustment rate, it is easier to obtain a higher attack success rate on the normally trained networks and in the black-box models. These results provide useful suggestions for constructing strong adversarial attacks in practice.
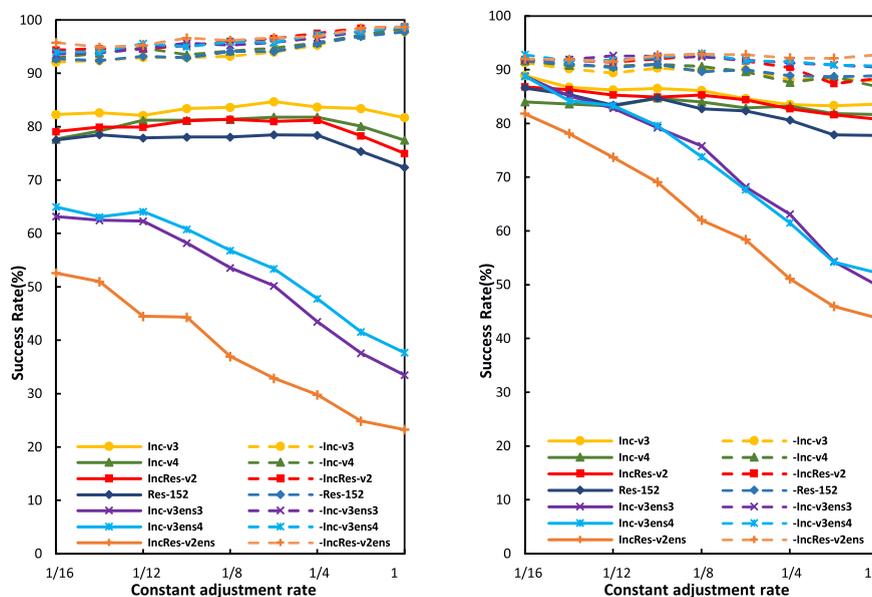
Fig. 6. Success rates of RT-MI-FGSM (left) and RT-DIM (right) under different constant adjustment rates *r*. Adversarial examples are generated on an ensemble of six networks and tested on the ensembled network (white-box setting, dashed line) and hold-out network (black-box setting, solid line).

5. Conclusions and future work

In this paper, we propose a new attack method based on data augmentation that randomly transforms the brightness of the input image at each iteration in the attack process to alleviate overfitting and generate adversarial examples with more transferability. Compared with traditional FGSM related methods, the results on the ImageNet dataset show that our proposed attack method has much higher success rates for black-box models, and maintains similar success rates for white-box models. In particular, our method is combined with DIM to form RT-DIM to further improve the success rates for black-box attacks on adversarially trained networks. Moreover, the method of attacking ensemble models is used simultaneously to further improve the transferability of adversarial examples. The results of this enhanced attack show that the average black-box attack success rate of RT-DIM on adversarially trained networks outperforms DIM by a large margin of 24.3%. Our work of RT-MI-FGSM suggests that other data augmentation methods may also be helpful to build strong attacks, which will be our future work, and the key is how to find effective data augmentation methods for iterative attacks. This inspires us to continue to explore the nature of adversarial examples, study the differences among data augmentation methods, and explore more

ways to improve model generalization performance. It is hoped that the proposed attack method can help evaluate the robustness of the models and the effectiveness of different defense methods and build deep learning models with higher security.

Declaration of Competing 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.

Authors' contributions

Bo Yang and Hengwei Zhang contributed equally to this work.

Acknowledgments

References

[1] I.S. Krizhevsky and G.E. Hinton, ImageNet classification with deep convolutional neural networks," *Commun. ACM* **60**(6) (2017), 84–90.

[2] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations*, ICLR (2015).

[3] C. Szegedy, et al., Going deeper with convolutions, in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Boston, MA, USA, (2015), 1–9.

[4] K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, (2016), 770–778.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I.J. Goodfellow and R. Fergus, Intriguing properties of neural networks, In *Yoshua Bengio and Yann LeCun, editors, 2nd International Conference on Learning Representations*, ICLR (2014).

[6] I.J. Goodfellow, J. Shlens and C. Szegedy, Explaining and harnessing adversarial examples, In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations*, ICLR (2015).

[7] K. Eykholt, et al., Robust physical-world attacks on deep learning visual classification, in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, (2018), 1625–1634.

[8] Y. Liu, X. Chen, C. Liu and D. Song, Delving into transferable adversarial examples and black-box attacks, In *5th International Conference on Learning Representations*, ICLR (2017).

[9] A. Kurakin, I. Goodfellow and S. Bengio, Adversarial examples in the physical world, (2016), *arXiv: 1607.02533*. [Online]. Available: https://arxiv.org/abs/1607.02533.

[10] Y. Dong, et al., Boosting adversarial attacks with momentum, in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, (2018), 9185–9193.

[11] D.P. Kingma and J. Ba, Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations*, ICLR (2015).

[12] I. Sutskever, J. Martens, G. Dahl and G. Hinton, On the importance of initialization and momentum in deep learning, in *Proc. Int. Conf. Mach. Learn.*, Atlanta, GA, USA, (2013), 1139–1147.

[13] T. Tieleman and G. Hinton, Lecture 6.5 - RMSProp, coursera: Neural networks for machine learning, *Tech Rep* (2012).

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, *International Journal of Computer Vision* **115**(3) (2015), 211–252.

[15] C. Xie, et al., Improving transferability of adversarial examples with input diversity, in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Long Beach, CA, USA, (2019), 2725–2734.

[16] B. Biggio, et al., Evasion attacks against machine learning at test time, in *Proc. Mach. Learn. Knowl. Discovery Databases*, Berlin, Heidelberg, (20130, 3876–402.

[17] Y. Zhang, G. Pu, M. Zhang and Y. William, Generating adversarial examples for DNN using pooling layers, *J Intell Fuzzy Syst* **37** (2019), 4615–4620.

[18] Y. Dong, T. Pang, H. Su and J. Zhu, Evading defenses to transferable adversarial examples by translation-invariant attacks, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), 4312–4321.

[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, Towards deep learning models resistant to adversarial attacks, In *6th International Conference on Learning Representations* ICLR (2018).

[20] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in *Proc. IEEE Symp. Secur. Privacy*, San Jose, CA, USA, (2016), 582–597.

[21] C. Xie, J. Wang, Z. Zhang, Z. Ren and A.Y. uille, Mitigating adversarial effects through randomization, In *International Conference on Learning Representations*, (2018).

[22] C. Guo, M. Rana, M. Cisse and L. van der Maaten, Countering adversarial images using input transformations, In *6th International Conference on Learning Representations*, ICLR (2018).

[23] P. Samangouei, M. Kabkab and R. Chellappa, Defensegan: Protecting classifiers against adversarial attacks using generative models, (2018), *arXiv: 1805.06605*. [Online]. Available: https://arxiv.org/abs/1805.06605.

[24] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang and W. Wen, Feature distillation: Dnn-oriented jpeg compression against adversarial examples, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), 860–868.

[25] C. Mao, et al., Multitask Learning Strengthens Adversarial Robustness, (20200, *arXiv: 2007.07236*. [Online]. Available: https://arxiv.org/abs/2007.07236.

[26] J.M. Cohen, E. Rosenfeld and J. Zico Kolter, Certified adversarial robustness via randomized smoothing, In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML* (2019), 1310–1320.

[27] A. Kurakin, I. Goodfellow and S. Bengio, Adversarial machine learning at scale, (2016), *arXiv: 1611.01236*. [Online]. Available: https://arxiv.org/abs/1611.01236.

[28] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh and P. McDaniel, Ensemble adversarial training: Attacks and defenses, In *6th International Conference on Learning Representations, ICLR* (2018).

[29] S. Liu, M. Shao and X.-P. Liu, GAN-based classifier protection against adversarial attacks, *J Intell Fuzzy Syst* **39** (2020), 7085–7095.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the inception architecture for computer vision, in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, (2016), 2818–2826.

[31] C. Szegedy, S. Ioffe, V. Vanhoucke and A.A. Alemi, Inception-v4, inception-ResNet and the impact of residual connections on learning, *presented at the AAAI Conf. Artif. Intell.*, San Francisco, California, USA, (2017).

[32] K. He, X. Zhang, S. Ren and J. Sun, Identity mappings in deep residual networks, in *Proc. Comput. Vision – ECCV 2016, Cham*, (2016), 630–645.