Research on a Method of Defense Adversarial Samples for Target Detection Model of Driverless Cars

Ruzhi Xu
School of Control and Computer Engineering, North China Electric Power University, China
xuruzhi@ncepu.edu.cn

Min Li
School of Control and Computer Engineering, North China Electric Power University, China
120202227124@ncepu.edu.cn

Xin Yang
School of Control and Computer Engineering, North China Electric Power University, China
120202227271@ncepu.edu.cn

Dexin Liu
School of Control and Computer Engineering, North China Electric Power University, China
120202227197@ncepu.edu.cn

Dawei Chen
School of Control and Computer Engineering, North China Electric Power University, China
cnchendawei@163.com

Abstract: *The adversarial examples make the object detection model make a wrong judgment, which threatens the security of driverless cars. In this paper, by improving the Momentum Iterative Fast Gradient Sign Method (MI-FGSM), based on ensemble learning, combined with L_∞ perturbation and spatial transformation, a strong transferable black-box adversarial attack algorithm for object detection model of driverless cars is proposed. Through a large number of experiments on the nuScenes driverless dataset, it is proved that the adversarial attack algorithm proposed in this paper have strong transferability, and successfully make the mainstream object detection models such as FasterRcnn, Sum of Squared Difference (SSD), YOLOv3 make wrong judgments. Based on the adversarial attack algorithm proposed in this paper, the parametric noise injection with adversarial training is performed to generate a defense model with strong robustness. The defense model proposed in this paper significantly improves the robustness of the object detection model. It can effectively alleviate various adversarial attacks against the object detection model of driverless cars, and does not affect the accuracy of clean samples. This is of great significance for studying the application of object detection model of driverless cars in the real physical world.*

Keywords: *Driverless cars, object detection, adversarial examples, parametric noise injection, adversarial training.*

Received March 24, 2022; accepted February 19, 2023
https://doi.org/10.34028/iajit/20/5/6

1. Introduction

Deep Neural Networks (DNNs) have achieved impressive results in the field of object detection [17, 25, 26], but adversarial examples can easily fool the most advanced deep neural network models [10, 32]. Therefore, more and more people are paying attention to the reliability and security of DNNs.

A large number of gradient-based adversarial attack algorithms have been proposed. At the same time, some defense methods have been proposed to be effective against transfer-based black-box adversarial attacks [8, 9, 33, 36]. But, the adversarial examples are highly related to the discrimination regions of the generated model. It is difficult to transfer to defense models that rely on different discriminative regions.

This paper considers more stringent black-box scenario. As shown in Figure 1, from the unified perspective of model robustness and regularization [35], this paper further studies how to enhance the robustness of the object detection system by using the adversarial attack algorithm proposed in this paper. The main contri-butions in this paper can be summarized as follows:

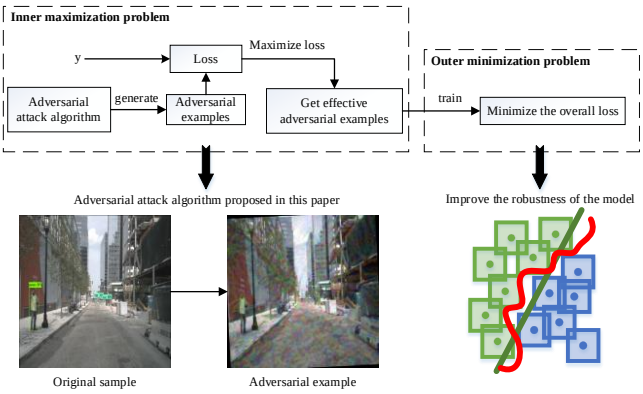

Figure 1. The process of algorithm.

The algorithm for generating adversarial examples in adversarial training is very important, so this paper proposed a more effective adversarial attack algorithm for object detection model of driverless cars, which uses a combination of perturbation and spatial transformation to generate more transferable

adversarial examples. Different from existing methods, this paper is aimed at the field of object detection. Moreover, based on DNNs selected by experiments, more effective adversarial examples are generated through integrated learning

1) Combined the adversarial example generation algorithm proposed in this paper (for adversarial training) with parametric noise injection, a defense model with strong robustness for object detection was generated.
2) A large number of experiments on the nuScenes driverless cars dataset [4] show that the adversarial attack algorithm proposed in this paper can effectively make typical object detection models misclassified, and the defense model proposed in this paper can effectively alleviate the threat of adversarial instances on the premise of ensuring the accuracy of clean samples. It is of great significance for researching the application of the object detection system of driverless cars in the real physical world.

2. Related Work

2.1. Adversarial Attacks

Many white-box algorithms have been proposed, such as Fast Gradient Sign Method (FGSM) [10], BIM (basic interative method) [13], Jacobian-based Saliency Map Attack (JSMA) [24], DeepFool [22], Carlini &Wagner's method (CW) [5], etc., Different from White-box attacks, the black-box attacks cannot obtain the parameters and gradients of the model, it is more consistent with the actual conditions of the attacker in the real physical world. A typical black-box attack is an iterative modification based on the label or label probability returned by the target model. Such as decision-based black-box attacks and score-based black-box attacks [6, 23, 30]. For example, single pixel attack and local search attack [23] are typical black-box attack algorithms, which construct the input based on the algorithm and then iteratively modify the input based on the model's feedback.

In addition, there are many open source DNNs available, such as GoogLeNet [20], ResNet [19], etc. Therefore, through the open source model, it is practical to construct a black-box attack algorithm based on the transferability of adversarial examples. For example, Liu *et al*. [18] proposed an ensemble learning method for adversarial examples. His method constructs transferable adversarial examples based on multiple DNNs, which shows that the adversarial examples generated by integrating multiple models are more transferable than the adversarial examples generated by a single model. Dong *et al*. [8] considered image translation on the basis of ensemble learning. Based on Convolutional Neural Network's (CNN's) translation-invariance to reduce the amount of calculation, they proposed black-box attack algorithms (TI-FGSM, TI-MI-FGSM, etc.,). This algorithm generates more transferable adversarial examples and can deceive the most advanced defense models with a high probability.

2.2. Defense of Object Detection Model

Object detection models (e.g., FasterRCNN [26], Sum of Squared Difference (SSD) [17], YOLOv3 [25], etc.,) are one of the most applied techniques for machine vision in the real physical world. Especially in the field of driverless cars, the use of a large number of object detection models is involved, such as traffic sign recognition [21], pedestrian collision warning, and automatic parking, etc.

Improving the robustness of DNNs to adversarial examples through adversarial training is an effective defense method that attracts researchers' attention [1, 32]. Buckman *et al*. [3] pointed out that, in the field of driverless cars, adversarial examples are not effective because the adversarial examples generated by adding perturbation are affected by the angle and direction. However, research by Athalye *et al*. [1] has shown that adversarial examples can reliably fool DNNs at different sizes and angles [27]. Not only that, the research of Engstrom *et al*. [9] also showed that even if the adversarial examples added perturbation are not carefully constructed, the adversarial examples generated only by the rotation and translation of the image may make the prediction of the DNN wrong. Although adversarial training cannot completely avoid all adversarial examples, the robustness of the model can be significantly improved by combining adversarial training. For example, the defense model proposed by Buckman *et al*. [3] and Samangouei *et al*. [27] further improved the defense ability of the model by combining adversarial training. In addition, Wang *et al*. [34] proposed a color-based feature representation method to accurately identify target vehicles in complex scenes. The method of noise injection has also been shown in recent researches to significantly improve the robustness of DNNs to adversarial examples [7, 35, 37].

3. Problem Statement

In the existing methods, black-box attacks that iteratively modify the returned results of the target model require a lot of access to the target model, which is not practical in the real physical world. In addition, the existing defense models are effective because they predict based on different discrimination regions, while the discrimination regions of normal training models are completely identical [2].

Therefore, the black-box attack algorithm proposed in this paper is mainly based on the transferability of adversarial examples, which generate adversarial examples through alternative models. Compared with

existing methods [2, 18], the method of this paper is aimed at target detection, and also generates adversarial examples through ensemble learning methods. In addition, this paper considers more complex image space transformation (rotation, translation), which has practical significance for the application of object detection models in the real physical world. And, from the perspective of model robustness and regularization, this paper proposes a defense method that combines parametric noise injection with adversarial training to significantly improve the robustness of object detection models to adversarial examples.

4. Robust Object Detection Model

In adversarial training, the algorithm for generating adversarial examples is very important, so this paper first proposes a black-box adversarial attack algorithm for object detection systems of driverless cars. Based on this algorithm, it will generate more transferable adversarial examples. In the algorithm, x is the original sample and y is the prediction label given by the model. For a classifier $f(x): x \to y$, it outputs the corresponding prediction label y according to the input sample x. The object detection model will output the bounding boxes and prediction probabilities. x^{adv} denotes the adversarial example generated based on the attack algorithm proposed in this paper. It is almost impossible for the human to distinguish between x and x^{adv}, but it can successfully deceive the object detection model (i.e., $f(x^{adv}) \neq y$). The goal of generating adversarial example is to make the loss function of the classifier (i.e., $L(x^{adv}, y)$) as large as possible, where L is the loss function.

Based on the above, combined parametric noise injection with adversarial training, it was experimentally determined that noise injection on the weights would have the best defense effect. Through the black-box adversarial attack algorithm and the defense method combining parametric noise injection with adversarial training proposed in this paper, the defense ability of the object detection model against adversarial examples is significantly improved. And experiments show that the defense model in this paper will not affect the accuracy of clean samples.

4.1. The Black-Box Attack with Strong Transfer-Ability

For stringent black-box scenario: the parameters and gradients of the target model cannot be obtained, and the model's feedback information cannot be obtained, but for open source DNNs, its parameters and gradients can be obtained. Based on the white-box attack algorithm, this paper expects to use a variety of open source DNNs to generate transferable adversarial examples that are effective for the target model. The white-box attack algorithms are mostly based on the calculation of gradients, which can generate adversarial examples with minimal perturbation. For example, the C&W [5] is a powerful white-box algorithm for generating adversarial examples. It expects to find the adversarial examples with the least perturbation, but the adversarial examples generated by C&W are not very transferable. Therefore, this paper considers Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [31], which can generate strong transferable adversarial examples. Fast Gradient Sign Method (FGSM) [10] is the earliest white-box attack algorithm to generate adversarial examples, as follows:

$$x^{adv} = x + \epsilon \cdot sign(\nabla_x L(x^{adv}, y)) \tag{1}$$

$\nabla_x L$ represents the gradient of the loss function for x. $sign(\cdot)$ is the sign function and ϵ represents L_∞ perturbation. The MI-FGSM is improved based on the FGSM. The transferability of the adversarial example is improved by introducing a momentum term, the formula is as follows:

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

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(g_{t+1}) \tag{3}$$

μ represents the decay factor. g_t represents the gradient information obtained in the t-th iteration and α represents the step size. The black-box attack algorithm proposed in this paper is based on MI-FGSM to generate adversarial examples that are more transferable and more in line with the practical application of object detection models.

With the development of the research on the security of DNN, a large number of methods that can defend the adversarial examples have been proposed [8, 9, 33, 36]. However, the reason why these defense methods can defend the transferable adversarial attack [12, 14, 18] is that when the transferable adversarial examples are generated based on the open source DNNs, the generated examples are highly related to the discriminative regions of these normal models, but it is difficult to transfer to the defense models that depend on different discriminative regions. So this paper expects to generate more transferable adversarial examples.

The method proposed in [2] successfully generated adversarial examples that made these defense models [8, 9, 33, 36] invalid by translating the image. Therefore, this paper considers the generation of adversarial examples with strong transferability to the object detection model by combining L_∞ perturbation and spatial transformation. Spatial transformation can not only improve the transferability of adversarial

examples, but also have practical significance for the use of object detection models. For example, the translation and rotation of images is equivalent to the change in the angle and distance between the camera and the object of the driverless cars.

The objective function that this paper expects to optimize is as follows:

$$\arg\max_{x^{adv}} L(T_{\delta u,\delta v,\theta}(x^{adv}), y), \quad \text{s.t.} \ \left\| x^{adv} - x \right\|_{\infty} \le \epsilon \tag{4}$$

L_{∞} is used to measure the gap between the adversarial examples and the original samples, and ϵ is the threshold. $T_{\delta u,\delta v}(x)$ means to translate the two-dimensional coordinates of the image x by δu , δv pixels respectively, $\delta u, \delta v \in \{-k,...,0,...,k\}$. $T_{\theta}(x)$ means to rotate the image x by θ degrees, $\theta \in \{-a,...,a\}$. When Equation (4) is optimized based on the MI-FGSM, for each original image x, the gradient of $(2k+1)\cdot(2k+1)\cdot 2a$ images needs to be calculated. For example, when k=10 and a=30, the gradient of 26460 images needs to be calculated. In order to reduce the computational overhead, this paper defines the objective function as follows:

$$\arg\max_{x^{adv}} L(x^{adv}, y), \quad \text{for} \ x^{adv} = T_{\delta u,\delta v,\theta}(x) \tag{5}$$

It is feasible to reduce the computational overhead based on the CNN's translation-invariance [2]. CNN has translation-invariance when the range of translation is small. When calculating the gradient:

$$\nabla_x L(T_{\delta u,\delta v}(x^{adv}), y) \approx \nabla_x L(x^{adv}, y) \tag{6}$$

Based on the above assumptions, the computational overhead is reduced: $(2k+1)\cdot(2k+1)\cdot 2a \to 2a$.

For the rotation of the image, assuming that point (0,0) is the center of the image, the formula is as follows:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{7}$$

For generating more transferable adversarial examples, it is impossible to use the first-order optimization method to obtain the optimal rotation angle θ because this is a non-convex optimization problem [30]. Therefore, this paper uses the random sampling method to randomly selects several angles in the range of $\{-a,...,a\}$, and then selects the angle that makes the model perform worst, which also reduces the computational overhead further: $(2k+1)\cdot(2k+1)\cdot 2a \to 2a \to$ several images.

Finally, the algorithm for generating black-box adversarial examples based on transferability is as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x L(x_t^{adv}, y)}{\left\| \nabla_x L(x_t^{adv}, y) \right\|_1} \tag{8}$$

$$x_{t+1}^{adv} = T_{\delta u,\delta v,\theta}(x_t^{adv}) + \alpha \cdot sign(g_{t+1}) \tag{9}$$

4.2. The Defense Model with Strong Robustness for Object Detection

Parametric noise injection can improve the robustness of the model. Based on the black-box adversarial attack algorithm proposed in this paper, combined parametric noise injection with adversarial training will further improve the robustness of the object detection model. The location of the parametric noise injection will seriously affect the defensive performance of the model against adversarial examples. He *et al*. [11] conducted experiments on the CIFAR-10 and MNIST [31] datasets using different structures of the ResNet network [19]. It shows that injecting noise on the weights has the best defense effect [29]. This paper conducted experiments on the nuScenes driverless cars dataset [28], which also proved that the defense effect of noise injection on weights is significantly better than noise injection on the input layer or activation layer. Moreover, the method of noise injection on weights is different from Differential Privacy (DP) [15]. The method of DP for noise injection is to improve the robustness of the model by sacrificing the accuracy of clean samples. However, in the field of driverless cars, the defense method of sacrificing the accuracy of clean samples is not applicable.

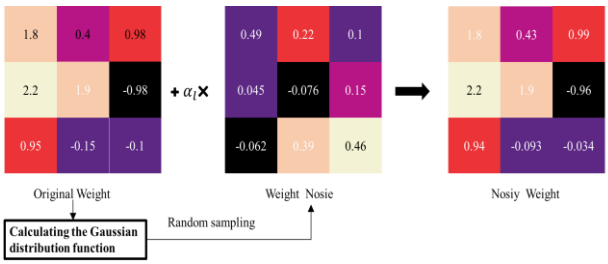

Figure 2. The flowchart of noise injection on the weight.

In this paper, the method of noise injection on weights combined with adversarial training not only improves the robustness of the object detection model, but also does not affect the accuracy of clean samples. The formula for noise injection on weights is as follows:

$$\tilde{w}_{l,i} = f(w_{l,i}) = w_{l,i} + \alpha_l \cdot n_{l,i} \tag{10}$$

$w_{l,i}$ is the noise-free weight value of the l layer of the neural network. $n_{l,i}$ is the noise value added to the corresponding weight, and α_l is the noise scaling

coefficient. $n_{l,i}$ is obtained by random sampling, and the sampling function is related to the Gaussian distribution of the weight values. α_l is used to control the size of the added noise value. As a parameter, it is optimized by back propagation. The process of noise injection on the weight of the 3×3 fully-connected layer is shown in Figure 2.

In Figure 2, the Gaussian distribution function of the original weight value is first calculated as the sampling function of the noise, and then the noise value is obtained by random sampling, in which the noise scaling coefficient a_l controls the size of the noise. Finally, the noise is injected into the original weight.

Based on the above, the combination of noise injection on the weight and adversarial training is shown in Equation (11):

$$\arg\min_{w}\{\arg\max_{x'\in P_\epsilon(x)} L(g(x^{adv}; f(w)), y\} \quad (11)$$

In the above, w is the weight to be optimized. X is the input image. Y is the target label. $P_\epsilon(x)$ is the dataset after adding perturbation, and its threshold is less than ϵ. L represents the loss function. The function f is a noise injection function, and $f(w)$ indicates that noise is injected into the weights. The function g is a function for generating adversarial examples, and $g(x^{adv})$ represents generating adversarial examples x^{adv} based on the adversarial attack algorithm proposed in this paper. Through the noise injection with adversarial training, the robustness of the object detection model is significantly improved.

Moreover, this paper improves the loss function to avoid the impact on the accuracy of clean samples, as shown in Equation (12):

$$L = R_c \cdot L(g(x; f(w)), y) + R_p \cdot L(g(x^{adv}; f(w)), y) \quad (12)$$

R_c and R_p are the weights for clean data loss term and adversarial data loss term, respectively. This paper sets R_c= R_p=0.5. The improved loss function L makes the defense model in this paper robust and significantly alleviates the threat of adversarial examples, while avoiding the impact on the accuracy of clean samples.

5. Experiments

5.1. Experimental Environment and Dataset

The nuScenes driverless cars dataset [28] is selected as the experimental dataset. The data in the nuScenes dataset comes from multiple cameras on driverless cars and contains 1.4 million pictures. In this paper, 100,000 images of 1600×900 are selected as experimental samples, including key elements such as pedestrians, various types of cars, traffic signs, and traffic lights. The recognition results using mainstream object detection models are shown in Figure 3.

Figure 3. Recognition results of open source object detection model.

As shown in Figure 3, the mainstream object detection models (FasterRCNN [26], SSD [16], YOLOv3 [25]) selected in this paper all recognize clean samples in the nuScenes dataset with high confidence.

Regarding the hyperparameters of the adversarial attack algorithm in this paper, we set the maximum perturbation ϵ as 16 and the decay factor μ as 1.0. The number of iterations is 10, and the step size is $\alpha = 1.5$. In addition, GoogLeNet [20], ResNet [28] (ResNet-18/50/101/152), Incept-v3, and VGG-16 are selected as the basis for researching the adversarial example generation algorithm proposed in this paper.

All experiments in this paper are implemented using TensorFlow. And in order to avoid the randomness in the process of adversarial example generation and parametric noise injection with adversarial training, all experimental results are reported in the form of the average of multiple experiments.

5.2. Generation of Adversarial Examples

- *Adversarial examples generated based on multiple DNNs*: the adversarial attack algorithm proposed in this paper relies on the transferability of adversarial examples. In order to evaluate the impact of different network models on the algorithm, this paper tests the transferability of adversarial examples generated by different network models to different object detection models, as shown in Table 1.

In Table 1, the horizontal axis represents the open source network model A, and the vertical axis represents the attacked object detection model B. Each cell represents the proportion of the adversarial examples generated by model A that can be correctly recognize by model B.

As shown in Table 1, the adversarial examples generated by the ResNet are more transferable to the FasterRcnn and YOLOv3, and the adversarial examples generated by the ResNet-101 show the strongest transferability; The adversarial examples generated by the GoogLeNet have poor transferability to the three object detection models, of which the transferability to the YOLOv3 is relatively strong; The adversarial examples generated by the Incept-v3 are more transferable to the YOLOv3, but are less

transferable to the FasterRcnn and SDD; The adversarial examples generated by the VGG-16 show strong transferability to the FasterRcnn and SSD, and poor transferability to the YOLOv3. Although the adversarial examples generated by different models have different transferability to different object detection models, they generally show strong transferability.

From the above research, this paper chooses ResNet-101, GoogLeNet, Incept-v3, and VGG-16 as the basis for the ensemble learning of the adversarial attack algorithm proposed in this paper.

Table 1. The transferability of adversarial examples.

	ResNet-18	ResNet-50	ResNet-101	ResNet-152	GoogLeNet	Incept-v3	VGG-16
FasterRcnn	56%	52%	45%	49%	79%	58%	51%
SSD	69%	73%	76%	78%	85%	64%	44%
YOLOv3	65%	62%	56%	60%	76%	51%	77%

Table 2. The effect of spatial transformation.

	perturbation only	**perturbation + translation**	**perturbation + rotation**	**perturbation + translation + rotation**
FasterRcnn	42%	39%	40%	37%
SSD	41%	38%	38%	35%
YOLOv3	44%	40%	41%	37%

Table 3. Comparison of transferable black-box attack algorithms.

	Ensemble-based approach	**TI-FGSM**	**TI-MI-FGSM**	**Algorithm of this paper**
FasterRcnn	42%	40%	38%	37%
SSD	41%	38%	37%	35%
YOLOv3	44%	40%	39%	37%

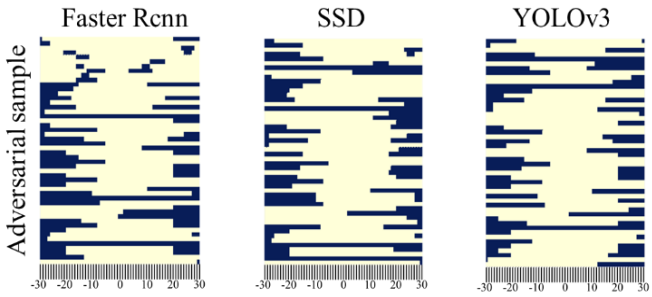

Figure 4. The effect of translation.

- *Effect of spatial transformation*: on the basis of ensemble learning, in order to further improve the transferability of the adversarial examples, this paper combines spatial transformation (i.e., the combination of adding perturbations and spatial transformation to generate strong transferable adversarial examples).

The spatial transformation includes translation and rotation. The effect of translation on generating adversarial examples is shown in Figure 4. In this paper, under the condition of the same perturbation and no rotation of the image, 50 translational adversarial examples are generated for each pixel in the range of $\{-10, 9,\ldots,0,\ldots, 9, 10\}$, and then the three object detection models are tested.

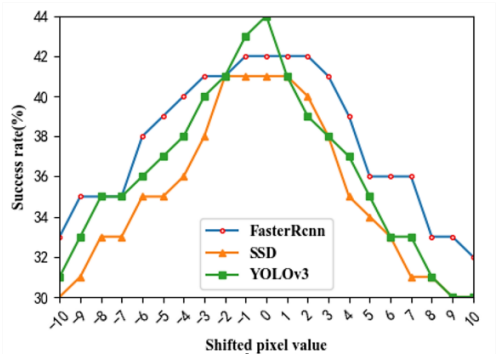

Figure 5. The effect of rotation.

In Figure 4, the horizontal axis represents the pixel value translated in each dimension, and the vertical axis represents the proportion that can be correctly recognized by the object detection model. As shown in Figure 4, the larger the pixel value of the image translation, the lower the accuracy of the model, which indicates that the image translation improves the transferability of the adversarial examples. Because this paper expects to use CNN's translation-invariance to significantly reduce the computational overhead, the image can only perform small translations: the pixel value translated in each dimension of the image does not exceed 10 pixels.

The effect of rotation on generating adversarial examples is shown in Figure 5. In this paper, under the condition of the same perturbation and no translation of the image, 50 rotational adversarial examples are randomly generated, and then rotated in the range of $\pm 30°$. Each generated adversarial example is tested under three object detection models.

In Figure 5, the horizontal axis represents the angle of rotation (positive numbers indicate clockwise rotation, negative numbers indicate counter clockwise rotation), the vertical axis represents the adversarial examples, and a small blue square represents an adversarial example that the object detection model cannot recognize.

As shown in Figure 5, the adversarial examples generated by only adding the L_∞ perturbation may be invalid, but the examples will become effective after rotating at a certain angle (i.e., the rotation and the addition of the L_∞ perturbation have an additive effect on the generation of the adversarial examples). Moreover, the small blue squares (i.e., adversarial examples) in the figure show similar distributions. On the one hand, it is shown that a larger rotation angle is more effective for improving the transferability of the adversarial example. On the other hand, it proves that the optimization of the rotation angle is a non-convex optimization problem. Therefore, this paper uses the random sampling method, randomly sampling in the range of $(-30° \sim +30°)$, and then choose the angle that makes the model perform the worst as the rotation

angle.

Table 2 shows the effect of spatial transformation on generating adversarial examples. The data in the table indicates the proportion of adversarial examples that can be correctly recognized by the object detection model.

It can be known from Table 2 that the combination of adding perturbation and spatial transformation improves the transferability of the adversarial examples.

- Adversarial *attack*: based on the above research, the adversarial attack algorithm proposed in this paper chooses ResNet-101, GoogLeNet, Incept-v3, and VGG-16 as the basis for the ensemble learning, and introduces spatial transformation on the basis of adding perturbation. In the algorithm, the image translation does not exceed 10 pixels, and the image rotation angle is randomly sampled in the range of $(-30° \sim +30°)$. The final generated adversarial examples are shown in Figure 6. The attack algorithm in this paper caused a slight distortion of the image, but in the real physical world, it is acceptable for object detection of driverless cars. And the main goal of this paper is to use the adversarial examples generated by the attack algorithm in this paper to train a strong robust object detection model.

Figure 6. Adversarial examples generated by the algorithm proposed in this paper.

Compared with other black-box attack algorithms, the adversarial examples generated by the algorithm proposed in this paper are more transferable to the object detection model, as shown in Table 3.

The data in Table 3 indicates the proportion of adversarial examples that can be correctly recognized by the object detection model.

5.3. Parametric Noise Injection with Adversarial Training

In order to alleviate the threat of adversarial examples, this paper combines parametric noise injection with adversarial training based on the proposed adversarial attack algorithm to improve the robustness of the object detection model.

- *The* effect *of noise injection location*: The location of the noise injection will significantly affect the robustness of the object detection model, so this paper compares the effects of noise injection at different locations, as shown in Table 4.

In Table 4, the horizontal axis represents different adversarial attack algorithms, and the vertical axis represents the object detection model after adversarial training with noise injection at different positions (-input: indicates that noise is injected at the input; -weight: indicates that noise is injected at the weight; -activation: indicates that noise is injected at the activation; -activation+weight: indicates that noise is injected at the activation and the weight). The data in Table 4 indicates the proportion of adversarial examples that can be correctly recognized by the object detection model. As shown in Table 4, noise injection on weights is most effective in improving the robustness of the object detection model.

- *Robustness test against adversarial attacks*: compared with vanilla adversarial training and training for noise injection on weights, the defense method based on this paper will significantly improve the robustness of the object detection model to adversarial examples, and will not affect the accuracy of clean samples, as shown in Table 5.

The data in Table 5 indicates the proportion of adversarial examples that can be correctly recognized by the object detection model. As shown in Table 5, vanilla adversarial training and training for noise injection on weights have poor defensive performance. Moreover, training for noise injection on weights will reduce the accuracy of clean samples. In contrast, the defense method proposed in this paper will further improve the robustness of the object detection model to adversarial examples without sacrificing the accuracy of clean samples

6. Conclusions

Considering the use of object detection system of driverless cars in the real physical world, this paper researches the influence of distance and angle on generating adversarial examples, and then proposes a black-box adversarial attack algorithm that combines L_∞ perturbation with spatial transformation

A large number of experiments have proved that the adversarial examples generated by the algorithm proposed in this paper have strong transferability and can successfully attack different object detection models. Then based on the adversarial attack algorithm in this paper, the method of noise injection on weights with adversarial training successfully improves the robustness of the object detection model to the adversarial examples, and the accuracy of clean samples was not affected.

Table 4. The effect of noise injection location (continued Table).

	Ensemble-based approach	**TI-FGSM**	**TI-MI-FGSM**	**Algorithm of this paper**
SSD-input	51%	50%	50%	47%
SSD-weight	**55%**	**53%**	**52%**	**50%**
SSD-activation	51%	47%	47%	46%
SSD-activation+weight	52%	52%	51%	48%
YOLOv3-input	54%	52%	51%	50%
YOLOv3-weight	**57%**	**56%**	**56%**	**54%**
YOLOv3-activation	52%	51%	51%	50%
YOLOv3-activation+weight	55%	53%	52%	51%

Table 5. The defensive performance.

Model	**Defense method**	**Clean**	**Adversarial attacks**		
			Ensemble-based approach	**TI-MI-FGSM**	**Algorithm of this paper**
FasterRcnn	Vanilla adversarial training	85%	47%	41%	40%
	Training for noise injection on weights	71%	45%	40%	39%
	Defense method in this paper	**85%**	**58%**	**56%**	**54%**
SSD	Vanilla adversarial training	81%	45%	40%	39%
	Training for noise injection on weights	69%	42%	39%	37%
	Defense method in this paper	**80%**	**55%**	**52%**	**50%**
YOLOv3	Vanilla adversarial training	84%	46%	40%	40%
	Training for noise injection on weights	72%	45%	40%	39%
	Defense method in this paper	**83%**	**57%**	**56%**	**54%**

References

[1] Athalye A., Engstrom L., Ilyas A., and Kwok K., "Synthesizing Robust Adversarial Examples," *arXiv Preprint*, 2017. https://doi.org/10.48550/arXiv.1707.07397

[2] Bietti A., Mialon G., Chen D., and Mairal J., "A Kernel Perspective for Regularizing Deep Neural Networks," *arXiv Preprint*, 2018. https://doi.org/10.48550/arXiv.1810.00363

[3] Buckman J., Roy A., Raffel C., and Goodfellow I., "Thermometer Encoding: One Hot Way to Resist Adversarial Examples," *in Proceedings of the International Conference on Learning Representation*, Vancouver, 2018.

[4] Caesar H., Bankiti V., Lang A H., Vora S., Liong V., Xu Q., Krishnan A., Pan Y., Balden G., and Beijbom O., "Nuscenes: A Multimodal Dataset for Autonomous Driving," *arxiv Preprint*, 2019.

[5] Carlini N. and Wagner D., "Towards Evaluating the Robustness of Neural Networks," *in Proceedings of the IEEE Symposium on Security and Privacy*, San Jose, pp. 39-57, 2017. doi: 10.1109/SP.2017.49.

[6] Chen P., Zhang H., Sharma Y., Yi J., and Hsieh C., "Zoo: Zeroth Order Optimization Based Black-Box Attacks to Deep Neural Networks without Training Substitute Models," *in Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, Texas, pp. 15-26, 2017. https://doi.org/10.1145/3128572.3140448

[7] Dong Y., Liao F., Pang T., Su H., and Zhu J., Hu X., and Li J., "Boosting Adversarial Attacks with Momentum," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, pp. 9185-9193, 2018. https://doi.org/10.48550/arXiv.1710.06081

[8] Dong Y., Pang T., Su H., and Zhu J., "Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CA, pp. 4312-4321, 2019. https://doi.org/10.48550/arXiv.1904.02884

[9] Engstrom L., Tran B., Tsipras D., Schmidt L., and Madry A., "A Rotation and a Translation Suffice: Fooling Cnns with Simple Transformations," *arXiv preprint*, 2017.

[10] Goodfellow I., Shlens J., and Szegedy C., "Explaining and Harnessing Adversarial Examples," *arXiv preprint* arXiv:1412.6572, 2014. https://doi.org/10.48550/arXiv.1412.6572

[11] He Z., Rakin A., Fan D., "Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness Against Adversarial Attack," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CA, pp. 588-597, 2019. https://doi.org/10.48550/arXiv.1811.09310

[12] Krizhevsky A., "Learning Multiple Layers of Features from Tiny Images," Technical Report,

University of Toronto, 2009.

[13] Kurakin A., Goodfellow I., and Bengio S., "Adversarial Examples in the Physical World," *arxiv Preprint*, 2016. https://doi.org/10.48550/arXiv.1607.02533

[14] LeCun Y., Bottou L., Bengio Y., and Haffner P., "Gradient-Based Learning Applied To Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324. 1998. DOI: 10.1109/5.726791

[15] Lecuyer M., Atlidakis V., Geambasu R., Hsu D., and Jana S., "Certified Robustness to Adversarial Examples with Differential Privacy," *in Proceeding of the IEEE Symposium on Security and Privacy*, CA, pp. 656-672, 2019. DOI: 10.1109/SP.2019.00044

[16] Liu W., Anguelov D., Erhan D., Szegedy C., and Reed S., "SSD: Single Shot Multibox Detector," *in Proceedings of the European Conference on Computer Vision*, Amsterdam, pp. 21-37, 2016. https://doi.org/10.1007/978-3-319-46448-0_2

[17] Liu X., Cheng M., and Zhang H., and Hsieh C., "Towards Robust Neural Networks via Random Self-Ensemble," *in Proceedings of the European Conference on Computer Vision*, Munich, pp. 369-385, 2018.

[18] Liu Y., Chen X., Liu C., and Song D., "Delving into Transferable Adversarial Examples and Black-Box Attacks," *Arxiv Preprint*, 2016.

[19] Lu J., Sibai H., Fabry E., and Forsyth D., "No Need To Worry About Adversarial Examples in Object Detection in Autonomous Vehicles," *Arxiv Preprint*, 2017. https://doi.org/10.48550/arXiv.1707.03501

[20] Madry A., Makelov A., Schmidt L., Dimitris., and Vladu A., "Towards Deep Learning Models Resistant To Adversarial Attacks," *arXiv preprint*, 2017. https://doi.org/10.48550/arXiv.1706.06083

[21] Ottom M. and Al-Omari A., "An Adaptive Traffic Lights System using Machine Learning," *The International Arab Journal of Information Technology*, vol. 20, no. 03, pp. 407- 418, 2023. https://doi.org/10.34028/iajit/20/3/13

[22] Moosavi-Dezfooli S., Fawzi A., Frossard P., "Deepfool: A Simple and Accurate Method To Fool Deep Neural Networks," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 2574-2582, 2016. https://doi.org/10.48550/arXiv.1511.04599

[23] Narodytska N. and Kasiviswanathan S., "Simple Black-Box Adversarial Perturbations for Deep Networks," *arXiv preprint*, 2016. https://doi.org/10.48550/arXiv.1612.06299

[24] Papernot N., McDaniel P., Jha S., Fredrikson M., and Celik Z., "The Limitations of Deep Learning in Adversarial Settings," *in Proceedings of IEEE European Symposium on Security and Privacy (EuroS&P)*, London, pp. 372-387, 2016. https://doi.org/10.48550/arXiv.1511.07528

[25] Redmon J. and Farhadi A., "Yolov3: An Incremental Improvement," *arXiv preprint*, 2018. https://doi.org/10.48550/arXiv.1804.02767

[26] Ren S., He K., Girshick R., and Sun J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems*, Montreal, pp. 91-99, 2015. https://doi.org/10.48550/arXiv.1506.01497

[27] Samangouei P., Kabkab M., Chellappa R., "Defense-Gan: Protecting Classifiers against Adversarial Attacks Using Generative Models," *arXiv preprint*, 2018. https://doi.org/10.48550/arXiv.1805.06605

[28] Sharif M., Bhagavatula S., Bauer L., and Reiter M., "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition," *in Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, Vienna, pp. 1528-1540, 2016. https://doi.org/10.1145/2976749.2978392

[29] Simonyan K. and Zisserman A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint* 2014. https://doi.org/10.48550/arXiv.1409.1556

[30] Szegedy C., Liu W., Jia Y., Sermanet P., and Reed S., "Going Deeper with Convolutions," *in Proceedings of the IEEE* conference *on Computer Vision and Pattern Recognition*, MA, pp. 1-9, 2015. https://doi.org/10.48550/arXiv.1409.4842

[31] Szegedy C., Vanhoucke V., Ioffe S., Shlens J., and Wonja Z., "Rethinking the Inception Architecture for Computer Vision," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 2818-2826, 2016. DOI: 10.1109/CVPR.2016.308.

[32] Szegedy C., Zaremba W., Sutskever I., Bruna J., Dumitru E., "Intriguing Properties of Neural Networks," *arXiv preprint*, 2013. https://doi.org/10.48550/arXiv.1312.6199

[33] Tramèr F., Kurakin A., Papernot N., Goodfellow I., and Boneh D., McDaniel P., "Ensemble Adversarial Training: Attacks and Defenses," *arXiv preprint*, 2017. https://doi.org/10.48550/arXiv.1705.07204

[34] Wang H., Dzulkifli M., and Azman I., "An Efficient Parameters Selection for Object Recognition Based Colour Features in Traffic Image Retrieval," *The International Arab Journal of Information Technology*, vol. 11, no. 3, pp. 308-314, 2014.

[35] Xiao C., Zhu J., Li B., He W., and Liu M., Song D., "Spatially Transformed Adversarial

Examples," *arXiv preprint*, 2018. https://doi.org/10.48550/arXiv.1801.02612

[36] Xie C., Wang J., Zhang Z., Ren Z., Yuille A., "Mitigating Adversarial Effects Through Randomization," *arXiv preprint*, 2017. https://doi.org/10.48550/arXiv.1711.01991

[37] Xie C., Zhang Z., Zhou Y., Bai S., and Wamg J., Ren Z., Yuille A., "Improving Transferability of Adversarial Examples with Input Diversity," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CA, pp. 2730-2739, 2019. https://doi.org/10.48550/arXiv.1803.06978

Ruzhi Xu, born in 1966, PhD, professor, her research interests include smart grid, AI security.

Min Li, born in 1997, MS candidate, his research interests include AI security, adversarial sample.

Xin Yang, born in 1997, MS candidate, her research interests include AI security, Differential privacy.

Dexin Liu, born in 1998, MS candidate, his research interests include AI security, 5G security.

Dawei Chen, born in 1995, MS candidate, his research interests include AI security, adversarial sample.