ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Enhancing Autonomous Driving: A Novel Approach of Mixed Attack and Physical Defense Strategies

Chuanxiang Bi[1], Shang Shi[2], and Jian Qu[3]*

[1] School of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, 11120, Thailand
[2] School of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, 11120, Thailand
[3] School of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, 11120, Thailand
* Corresponding E-mail: jianqu@pim.ac.th

Citation:
Bi, C.; Shi, S.; Qu, J. Enhancing autonomous driving: A novel approach of mixed attack and physical defense strategies. *ASEAN J. Sci. Tech. Report.* **2025**, *28*(1), e254093. https://doi.org/10.55164/ajstr.v28i1.254093.

Article history:
Received: May 13, 2024
Revised: October 8, 2024
Accepted: October 17, 2024
Available online: December 14, 2024

Abstract: Adversarial attacks are a significant threat to autonomous driving safety, especially in the physical world where there is a prevalence of "sticker-paste" attacks on traffic signs. However, most of these attacks are single-category attacks with little interference effect. This paper builds an autonomous driving platform and conducts extensive experiments on five single-category attacks. Moreover, we proposed a new physical attack - a mixed attack consisting of different single-category physical attacks. The proposed method outperforms existing methods and can reduce the accuracy of traffic sign recognition of an autonomous driving platform by 38%. Furthermore, we proposed a new anti-jamming model for physical adversarial defense, CBAM-ResNet26 & CBAM-Alexnet, which improves an autonomous driving platform's traffic sign recognition accuracy to 63% under mixed attack. Finally, experiments were also conducted with datasets with different ratios of adversarial attack examples, and the experimental results showed that in adversarial training, the higher the ratio of adversarial examples, the higher the recognition accuracy. However, a too high ratio would reduce the accuracy of normal traffic signs. Finally, the optimal ratio for physical adversarial defense training is 1:2.

Keywords: Autonomous Driving, Physical Attack, Mixed Attack, Physical Defenses, Adversarial Training.

1. Introduction

Adversarial attacks [1] have gradually become a research topic that attracts attention today. They can be used to deceive machine learning models and then produce wrong predictions. Even deep neural networks (DNN) [2] that have succeeded in recent years are also subject to such threats. Therefore, many aspects are susceptible to adversarial attacks, such as autonomous driving [3-5], image classification [6, 7], natural language processing [8-10], etc., thus creating huge security risks. There are two different attack forms: In digital form, the attacker can directly input the input digital image into the DNN classifier, and adversarial examples can also be generated by generative adversarial networks (GAN) [11-13]. However, since the generation process is challenging to control, using GANs to perform an aqueous attack on a given test image is difficult. (2)

Physical form, where the DNN classifier only accepts input from the camera, and the attacker can only present adversarial images to the camera. Kurakin et al. [14] printed adversarial pictures on paper and took pictures with a camera to successfully deceive the Inception v3 model. Evtimov et al. [15] placed black and white stickers on traffic signs to fool the classifier. Athalye et al. [16] created a 3D-printed turtle mistaken for a rifle or game by a DNN classifier. These illustrate that physical attacks require large or unlimited perturbations [17]. In contrast, digital attacks, it is usually small perturbations, but small perturbations are too subtle to be captured by cameras in complex physical world environments, which shows that although the generation of adversarial samples in the digital world is similar to that in the physical world, the generation methods in the digital world are difficult to transfer and apply to the physical world directly. On the one hand, there is a lot of noise in the actual physical environment, which may cause perturbation produced by perturbing unobservable limits. On the other hand, in the physical world, an attacker cannot directly modify the input data of the model. Moreover, most digital attacks are white-box attacks [18, 19], while most adversarial physical attacks are black-box attacks[20-22], so it is more practical to study adversarial physical attacks. Take the recognition of traffic signs by self-driving cars as an example. Under normal circumstances, self-driving cars correctly recognize real stop signs. However, if an attacker can physically and robustly manipulate the traffic sign, the deep neural network may misjudge it as other actions. Well, this could have serious consequences. Therefore, this paper focuses on traffic sign classification, establishes an autonomous driving platform, and deeply studies the nuances of the physical world. Through comparative experiments on the effectiveness of several single categories of physical attacks, such as QR code sticker attacks, Colored bar, and graffiti attacks, this paper proposes a real and effective adversarial physical attack, a mixed attack, can pose a huge threat to self-driving cars in recognizing traffic signs, and proves that physical adversarial perturbations can exist robustly under realistic assumptions.

Current research on adversarial attacks in the physical world reveals potential threats and encourages researchers to explore effective methods to defend against adversarial attacks. Several approaches have been proposed, such as adversarial training [23], an adversarial sample-based generation and training strategy designed to improve the ability of machine learning models against attacks. During training, the model is exposed to adversarial examples with minor perturbations to force it to learn representations that are robust to these perturbations. These techniques do not rely on detecting adversarial examples but formulate models that perform equally well under adversarial and normal inputs. In terms of physical defense, this paper proposes a model CBAM-ResNet26 & CBAM-Alexnet with better anti-interference performance based on the improvement of ResNet18 for the currently popular sticker attacks and the mixed attack proposed in this paper. It also proposes a model based on the currently popular adversarial training method. The optimal proportion of the dataset for adversarial training of physical attacks. This paper aims to reduce the threat of physical attacks on autonomous cars in identifying traffic signs by proposing a new model with better anti-interference and improving physical defense methods. Our main contributions to this paper are:

1: An extensive comparison of single-category physical attacks.

2: This paper proposed a novel mixed attack. When compared to single-category physical attacks, our mixed attack is more effective.

3: This paper proposes a new anti-interference model, CBAM-ResNet26 & CBAM-Alexnet. It has been proven through experiments that it exhibits higher accuracy in the face of interference, which is better than the current ResNet18 and Alex Net.

4: In terms of physical defense training methods, this paper established different proportions of datasets containing adversarial examples, trained different models, and compared them. As the proportion of adversarial examples increases, the recognition accuracy of smart cars gradually increases. However, an excessively high ratio of adversarial examples will reduce the model's recognition accuracy of normal traffic signs. Finally, this paper determined that the optimal ratio of the physical adversarial training dataset is 1:2.

2. Literature Review

2.1 Digital Attacks

Existing research on adversarial examples mainly focuses on two domains: the digital world and the

physical world. Adversarial examples from the digital world are embedded directly into the input of the model, and their amplitude is usually limited by the lp norm, such as l∞-norm [6, 24, 25], l2-norm, and l0-norm, to ensure that they are overlooked. In a white-box attack, after the attacker knows all the model details, he can use the gradient to generate perturbations [26, 27]. For example, the L-BFGS attack [28] was an early method to deceive models in image recognition tasks such as deep neural networks. This attack could already successfully deceive state-of-the-art classification models at the time, such as Alex Net [6] and QuocNet [29], resulting in the Misclassification of many image instances. In addition, digital attacks may also be black-box attacks, where the attacker can only query the target model and obtain the corresponding output without understanding its internal structure. In a black-box scenario, an attacker can exploit the cross-model generalization capabilities of adversarial examples or reconstruct the internal information of the model through multiple queries. Papernot et al. [30] proposed the Substitute Black Box Attack (SBA), an early practical black box method. The key idea is to train a substitute model to imitate the target black box model and use the white box attack method on the substitute model.

2.2 Physical Attacks

Academic research in recent years has examined whether adversarial attacks are effective in the physical world. Some research has been applied to the physical world by generating attacks in the digital world. ShapeShifter [31] generates adversarial stop signs with complex patterns by executing formulas, but applying them to real traffic signs is difficult. AdvCam [32] uses neural style transfer to create naturally corroded stop signs similar to the surrounding environment, making it difficult for humans to notice and maintain significant disturbances. However, those perturbations that modify the image's pixels must survive printing to be effective, and there is a high likelihood that the pixels will be lost in the printer's copy. Also, noise-based perturbations are impractical in the physical world because they are difficult to capture with distant cameras. Lu et al. [33] conducted experiments using traffic signs printed on poster paper and added fixed perturbations to the background of the traffic signs. However, this method is less feasible in practical applications because, in real life, the background is often not fixed. Evtimov et al. [15] proposed a Robust Physical Perturbation (RP2) attack, successfully showing how the generated landmarks can be physically implemented and fool the victim model in a controlled environment. The sticker attack proposed in it is a scenario where the attacker prints some stickers containing adversarial perturbations and pastes them on real traffic signs. Still, the mask used in the sticker attack only allows the perturbation to be generated on certain parts of the traffic signs, and the attack target is more limited to stop signs. The mixed attack proposed in this paper fully considers the suspiciousness of the attack, and its suspiciousness is low; this is achieved by adding a small part of the modified object, and the modification is similar to the common "noise" in the physical world; for example, people who are outside all year round There will be some stains, black spots on the traffic signs, advertisements posted on the traffic signs, etc. Therefore, to most observers, a mixed attack on the traffic signs may only be seen as vandalism or lack of quality and will not Arouse too much suspicion. Moreover, the mixed attack in this paper can attack a wide range of traffic sign types in terms of attack targets and can have different combinations of attack methods.

2.3 Physical Defenses

Adversarial training [23] is an effective defense method that retrains the model by injecting adversarial examples into the training set at each training iteration. Although adversarial training is relatively simple to implement and performs reasonably well, an adversarial trained model may still be susceptible to other types of adversarial examples that the model was not trained on. For example, adversarial training using examples generated by single-step gradient-based attacks (e.g., FGSM) shows robustness against the same type of attacks even when evaluated at scale, but not against single-step methods. The training results in less robust models for iterative gradient-based attacks such as BIM. Many methods have been proposed to make models more robust to adversarial attacks. Ensemble adversarial training [34] is a variant of adversarial training in which the model is retrained on generated adversarial examples to attack other pre-trained models. This decoupling of the target model and adversarial training examples overcomes the overfitting problem observed in ordinary versions of adversarial training. Trammell et al. [35] argue that ensemble adversarial

training is a good approximation of internal maximization due to the transferability of adversarial examples between different models. Furthermore, since the adversarial example generation process is independent of the trained model, they hypothesize that overall adversarial training will be more robust to future black-box attacks than standard adversarial training. Chen et al. [36] proposed the Drift Diffusion Model (DDDM) based on Dropout. They applied dropout technology to generate a cropped model of the target model and then used the output of each cropped model to perform threshold accumulation. When the decision evidence accumulates to a certain threshold, the result is a robust output, formulating a model that performs equally well under adversarial and normal inputs. Essentially, many of these methods attempt to reduce the model's sensitivity to irrelevant changes in the input, effectively regularizing the model to reduce the attack surface and limit the response to manifold perturbations. However, this paper starts from the model itself, proposes a new anti-interference model, and applies it to physical adversarial training based on the principles of adversarial training methods. By studying the different proportions of adversarial examples in the dataset, the results of physical adversarial training are obtained the optimal proportion of the dataset.

3. Materials and Methods

3.1 Mixed Attacks

This paper proposes an innovative adversarial physical attack strategy—mixed attack. A mixed attack is a multi-category combination attack, as shown in Figure 1. It mainly consists of QR code sticker attacks, Colord bars, graffiti attacks, and several single-category physical jamming combinations, and these attack elements can be combined in different ways or pasted in different locations to carry out mixed attacks on traffic signs, so according to the combination of the way and the location of the paste, the mixed attacks are very varied styles. Therefore, many types of mixed attacks depend on the combination and location. These single-category interferences are relatively common in the physical world. In subsequent experiments, we will compare the effectiveness of physical attacks such as QR code sticker attacks, Colored bars, and graffiti attacks. Compared with traditional single-category physical attacks, mixed attacks combine multiple elements and are more likely to mislead the judgment of deep learning models. The attack method mainly uses physical stickers. As shown in Figure 2, the mixed attack is carried out by sticking it on the traffic sign to be interfered with. It is designed to confuse the traffic sign recognition system of the autonomous car, causing it to generate wrong predictions.

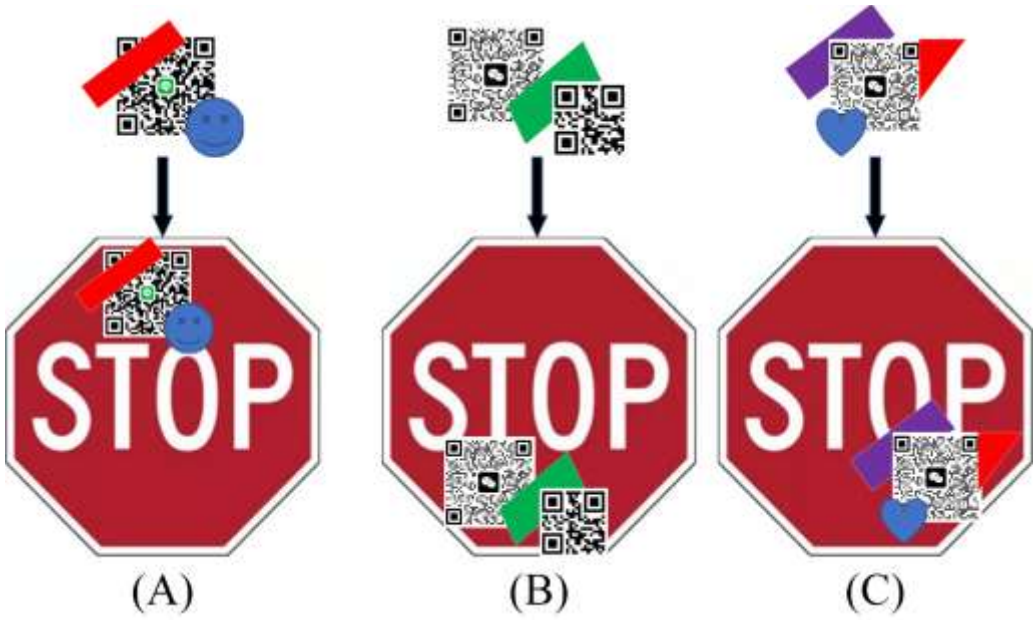

Figure 1. Example of mixed attack applied to a traffic traffic sign. (A) is a mixture of three attacks: QR code sticker attack, Colord bar, graffiti attack pasted on a traffic sign, (B) is a mixture of QR code sticker attack and Colord bar pasted on a traffic sign, (C) is a mixture of Colord bar, graffiti attack three

attacks mixed pasted on the traffic sign but at different locations on the traffic sign.

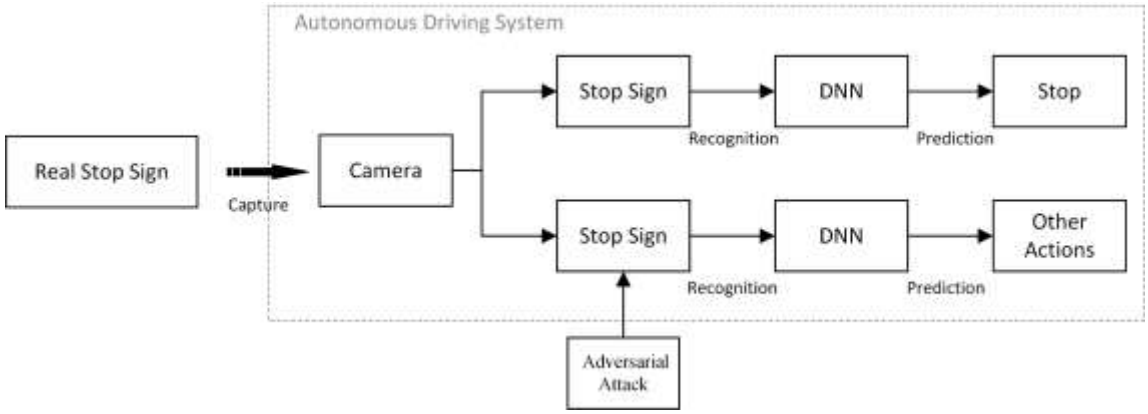

Figure 2. In the face of a real stop sign, the self-driving car normally acquires an image through the camera and then makes a stop action. Still, in the case of an adversarial attack, the camera acquires an image of the attack and is interfered with to make other actions.

3.2 CBAM-ResNet26&CBAM-AlexNet

We introduce the CBAM (Convolutional Block Attention Module) module, as shown in Figure 3, which combines channel and spatial attention mechanisms to enhance the model's feature extraction and learning capabilities. Channel attention obtains global information through global average pooling and maximum pooling, learns correlations between channels using convolutional layers and activation functions, and highlights essential channels by applying weights to the feature map through a sigmoid function. The spatial attention module, on the other hand, captures the spatial information between channels by performing average and maximum pooling along the channel dimensions and applies the attention weights to optimize the feature map. This channel and spatial attention combination enables the model to adaptively adjust the importance of different channels and locations in the feature map, thus improving the model's ability to recognize different targets and further enhancing the overall performance.

Based on this, we propose CBAM-Resnet26 & Alexnet. ResNet, as a deep convolutional neural network, solves the problem of gradient vanishing and explosion during deep network training by the design of residual blocks. Each residual block contains two consecutive convolutional layers, and the output of the second convolutional layer is added to the first output to form a residual connection. Based on the original ResNet-18 model, we propose a deeper Resnet26 model by adding convolutional layers and batch normalization layers to enhance its representation and feature learning capabilities. Subsequently, the CBAM module is introduced to construct CBAM-Resnet26. For better performance comparison, we added the CBAM module to the classical model AlexNet to form the CBAM-AlexNet model.

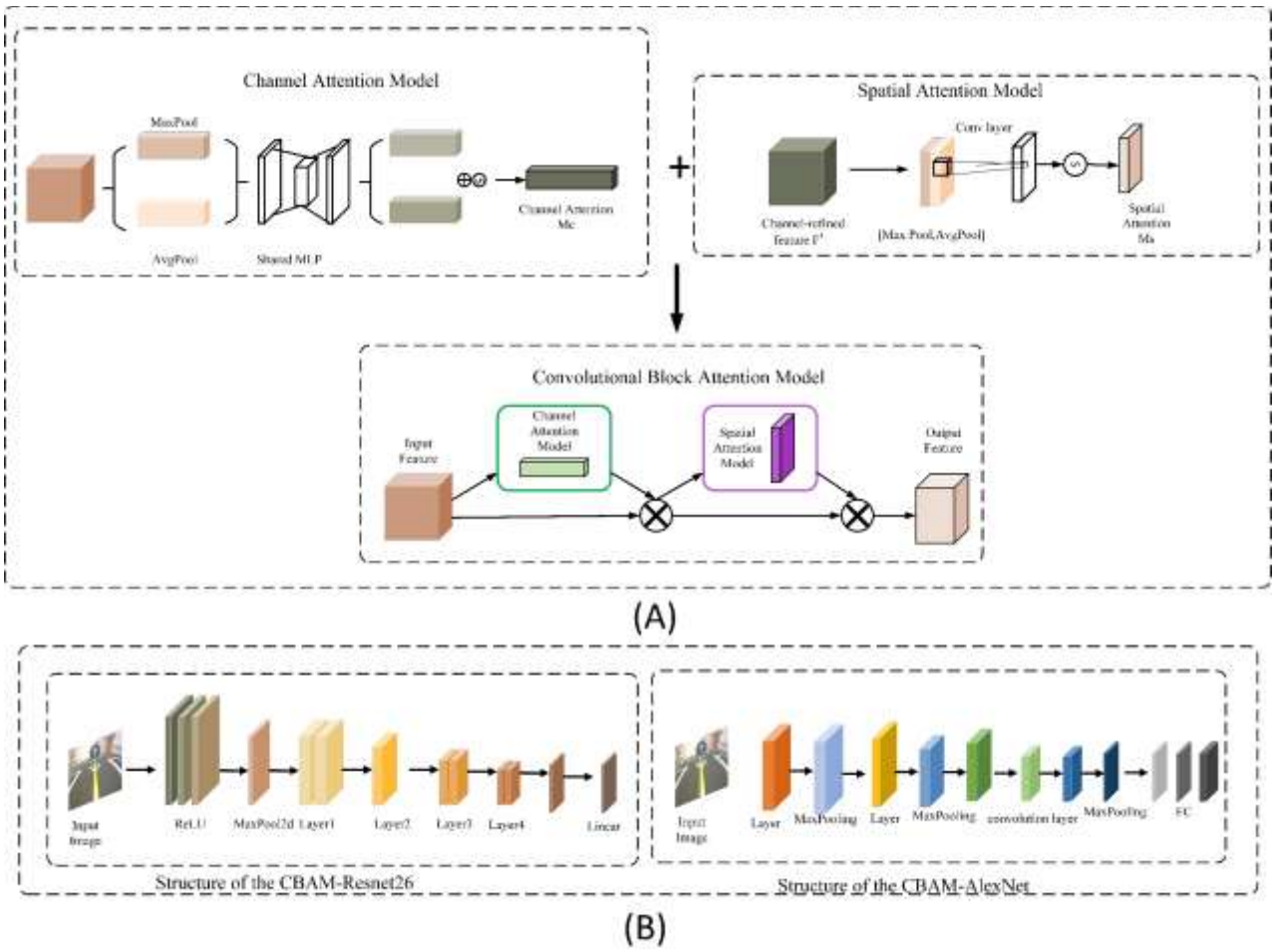

Figure 3. (A) CBAM consists of a combination of channel attention and spatial attention. (B) CBAM-ResNet26 and CBAM-AlexNet framework diagram proposes a deeper ResNet26 model based on the ResNet18 model.

3.3 Adversarial Physics Training

In addition to proposing a new model with better anti-interference performance to improve the robustness of the autonomous driving system to physical attacks, this paper also introduces an adversarial training method. It explores the optimal proportion of adding adversarial physical examples to the dataset. Adversarial physics training improves the robustness of the model by adding adversarial examples to the training data. We collected the adversarial examples in this paper, and then we constructed a training dataset that contains images of normal traffic signs and images of traffic signs using mixed attacks. However, the ratio of normal and adversarial example images in the dataset here is not optimal. For specific research, we tried datasets with different proportions of adversarial examples to train new model test effects and obtain the optimal proportion of adversarial examples in adversarial physics training. Specifically, we set different proportions of attack samples, such as 20%, 30%, 40%, 50%, etc., and trained corresponding deep-learning models. We use advanced deep learning models to train the model during the training process. During the experiment, we conducted a mixed attack on traffic signs. We evaluated the training effect of the model at different scales and recorded its performance.

4. Experimental Setup

4.1 Experimental environment

This research is not limited to only using deep neural network (DNN) classifiers for testing. This

research built an autonomous driving platform to study autonomous cars traffic sign recognition problems. It used a smart car equipped with a Jetson Nano motherboard as the research object to simulate real scenarios. Jetson Nano is a high-performance computing platform specially designed for embedded systems. It has the characteristics of small size and low power consumption. It is very suitable for use in scenarios such as smart cars so that they can complete tasks related to autonomous driving. The smart car is shown in Figure 4. To be close to the most primitive self-driving car, the smart car does not add any sensors except a camera. It is also combined with a deep neural network classifier to observe the traffic sign recognition process in a more realistic situation; thus, experiments were conducted to evaluate model performance more accurately.

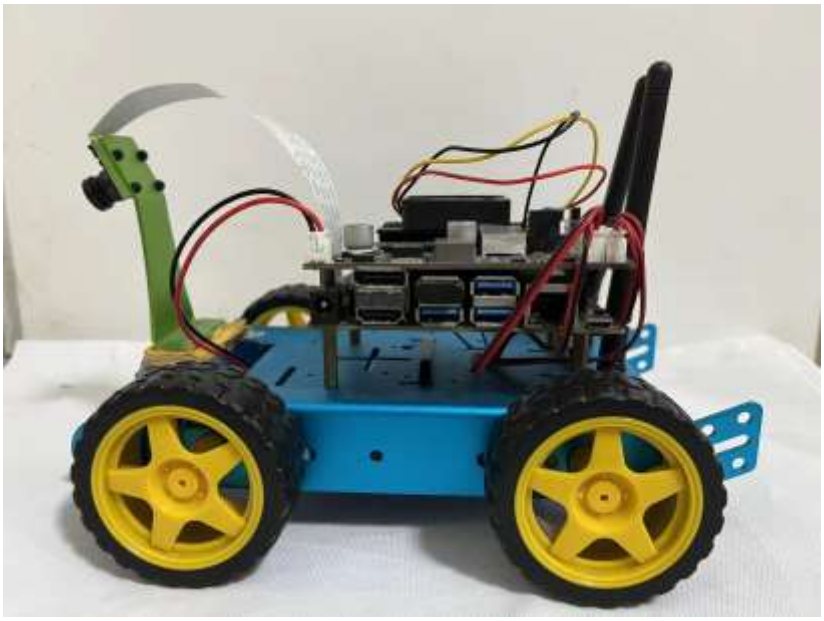

Figure 4. Smart car structure.

In the experimental environment, since the smart car continues to move forward while driving, the distance and angle of its camera relative to the traffic sign will change, so we need to fully consider the influence of these factors when designing the experiment. To simulate the real road environment, we chose a map to imitate the black road as the road where the smart car travels, as shown in Figure 5. At the same time, we took a series of measures to ensure the stability of the experimental environment. To create consistent light conditions, we blocked out external light, minimized other potentially intrusive noises, and kept the interior lighting steady, with a light intensity of 170lux. This paper chooses the four most basic functional indicators of traffic signs in the physical world: “forward,” “left,” “right,” and “stop.” To prevent the experiment from being accidental, two common traffic signs were chosen for each type of functional signage; these eight traffic signs had a radius of 12 cm, and we made sure that they had bases made of the same material with a base height of 6 cm, as shown in Figure 6. Such a design can reduce the possibility of the model identifying traffic signs by observing differences in other aspects of the traffic signs and improve the reliability and reproducibility of experimental results. At the same time, we also considered the importance of increasing the number and diversity of traffic signs. By introducing more types of traffic signs, we can more comprehensively evaluate the model's ability to recognize different traffic signs, thereby improving the rigor and practicality of the research. This study can conduct effective autonomous driving research in a more realistic road environment through the above comprehensive design..

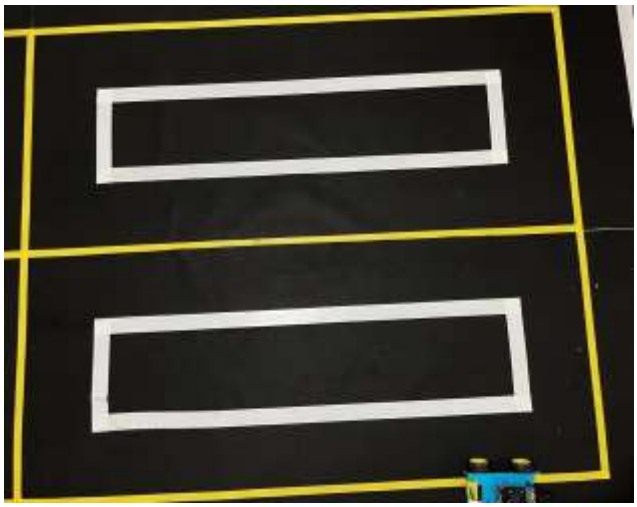

Figure 5. Smart car driving map.

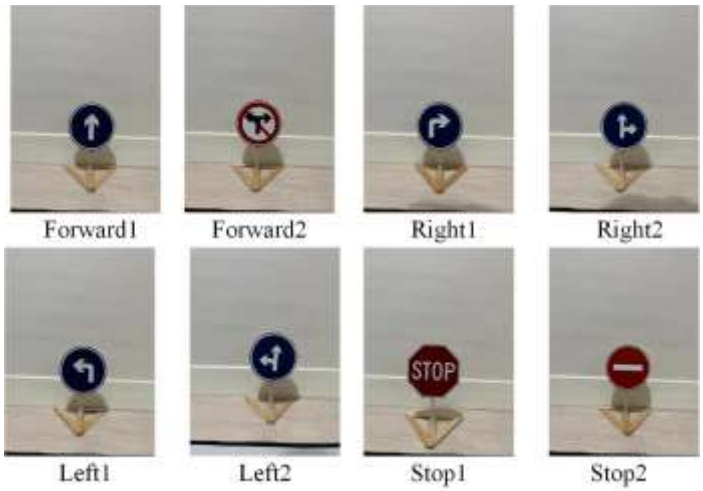

Figure 6. There are two types of "forward," "left," "right," and "stop" traffic signs, and there are a total of eight traffic signs.

4.2 Experimental design

In this paper's experimental design, mixed attacks apply multiple single-category interferences to traffic signs in a combined manner compared with the traditional single-category attack method. Therefore, mixed attacks have multiple combinations. We randomly add mixed attacks on the traffic signs, which can be grouped into 5 ways, as shown in Figure 7. A higher combination method (Figure 6) pastes these 5 ways of attack on 8 traffic signs in sequence. It conducts 10 tests at five different angles and distance settings (tests at each different angle and distance Twice). As shown in Figure 8, this multi-faceted testing method gives us a more comprehensive understanding of the model's ability to cope with complex environments. We recorded the smart car's response to the traffic signs during the test. We displayed the action probability distribution in real-time through the car's visual interface, as shown in Figure 9. When the probability value is higher than 0.5, the smart car will judge it as belonging to a category with a high probability value and successfully make the corresponding action. When the probabilities of the four categories do not exceed 0.5, the smart car will keep swinging left and right and pass the action. By analyzing probability distribution, we can more accurately assess the impact of attacks on the model. This study can conduct effective autonomous driving research in a more realistic road environment through the above comprehensive design.

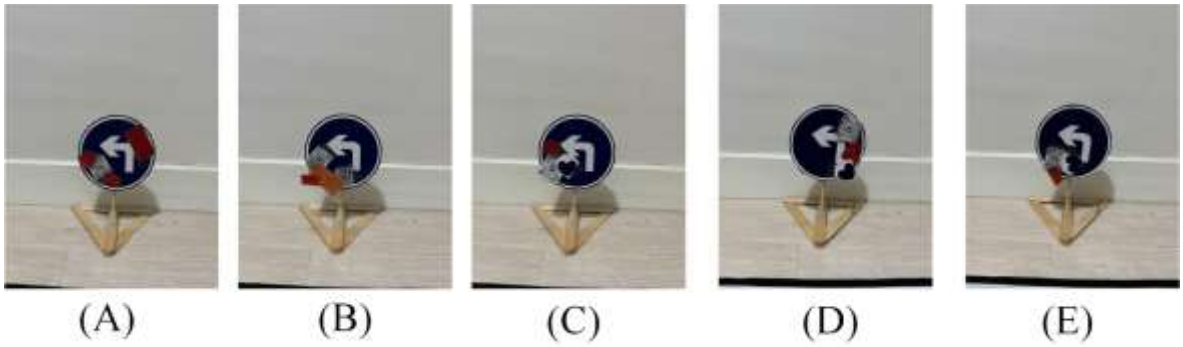

Figure 7. (A) QR code sticker attack, graffiti attack on the left side of the traffic sign, and Colord bar on the right side. (B) There were two QR code sticker attacks and a Colord bar on the left side of the traffic sign. (C) Only QR code sticker attacks and graffiti attacks are posted on the left side of the traffic sign. (D) Only QR code sticker attacks and graffiti attacks are posted on the right side of the traffic sign. (E) Only QR code sticker attacks, graffiti attacks, and Colord bars are posted below the traffic sign.

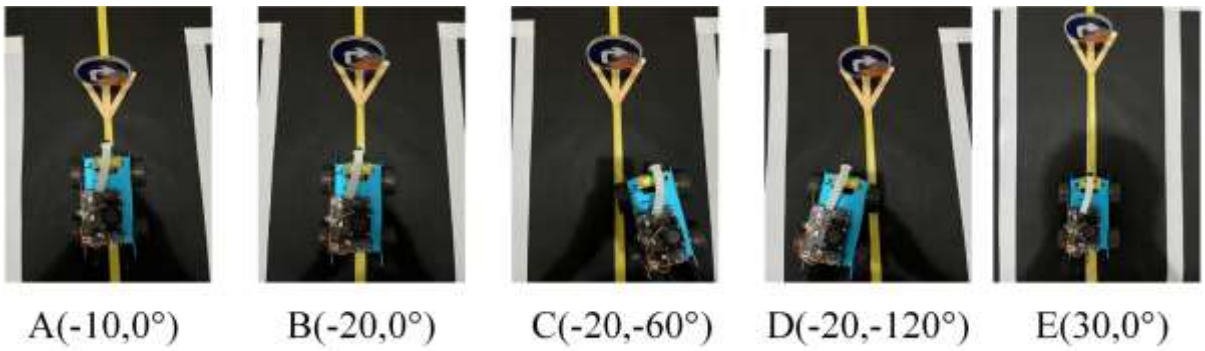

Figure 8. The location of the signpost is the origin (0, 0), the x-axis is the angle, and the y-axis is the distance. a is 10cm from the signpost, 0°, b is 20cm from the signpost, 0°, c is 20cm from the signpost, -60°, d is 20cm from the signpost, -120°, and e is 30cm from the signpost, 0°.

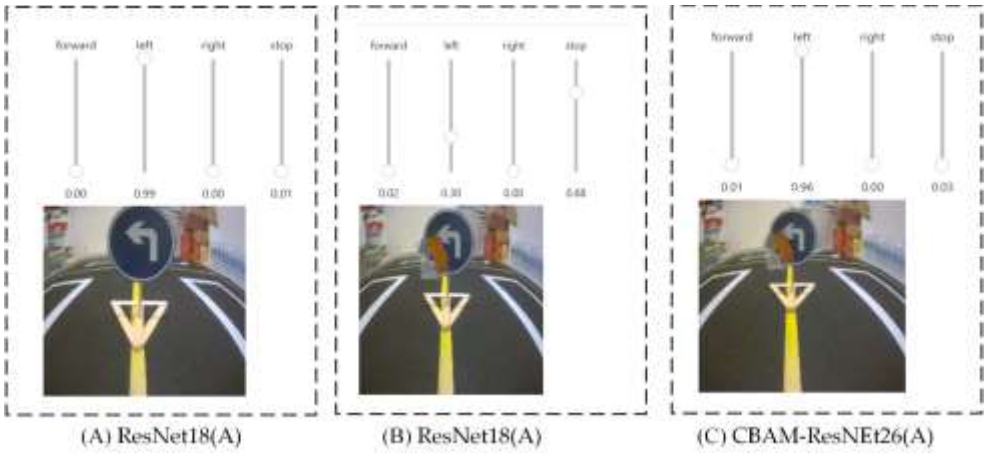

Figure 9. Probability distributions of actions shown in the visualization interface of the smart car, (A) ResNet18(A) test without physical attack, the probability of left is 0.99, (B) ResNet18(A) test with physical attack, the probability of left is 0.30, (C) CBAM-ResNet26(A) test with physical attack, the probability of left is 0.96.

4.3 Dataset

Aiming at the traffic sign recognition problem proposed above to study the traffic sign recognition

problem of the autonomous driving system, this study used self-collected data. Each dataset contains four categories: forward, left, right, and stop. As shown in Table 1, first, we established a basic dataset A, which collects normal pictures, which can ensure that smart cars can recognize traffic signs normally. Secondly, to deeply explore the effect of mixed attacks and the impact of the ratio of normal pictures to models in the dataset on model performance, we designed a series of adversarial physics training datasets, namely datasets B, C, D, and E, where each the datasets all include varying proportions of adversarial examples to evaluate the model's robustness and adversarial capabilities more fully. Next, we will detail the construction and design of these adversarial physics training datasets.

Table 1. The construction of five datasets.

Dataset	A	B	C	D	E
Number of attack images	0	800	1200	1600	2400
Normal images	4800	4000	3600	3200	2400
Total	4800	4800	4800	4800	4800
Proportions		1:5	1:3	1:2	1:1

4.3.1 Basic Dataset A

Firstly, we create a base dataset called Dataset A, designed to provide the basic data for the smart car to recognize traffic signs properly and make corresponding actions. The dataset contains four main categories: forward, left, right, and stop. We collected about 600 images covering eight types of traffic signposts and categorized them according to their function. There are about 1,200 images for each classification, totaling 4,800 images.

This basic dataset provides an important foundation for subsequent experiments, allowing smart cars to complete the basic functions of normal recognition of traffic signs. The model's recognition performance of traffic signs with different physical attacks can be compared on this basis. You can also compare the anti-interference effects of different models. Moreover, during the construction of Dataset A, we paid attention to the diversity and sufficiency of the data to ensure that the model can accurately identify and make decisions in various scenarios.

4.3.2 Dataset B

In Dataset B, we maintained a ratio of 1:5 between normal images and adversarial example images. Specifically, we collected 500 normal images for each traffic sign and 100 images with adversarial examples simultaneously. This ratio was chosen to enable the impact of the adversarial examples to be significantly represented in the dataset so that the robustness of the model can be better assessed.

4.3.3 Dataset C

In dataset C, we slightly reduced the proportion of adversarial examples, keeping the ratio of normal images to adversarial examples at 1:3. 450 normal images and 150 images with adversarial examples were collected for each traffic sign. By adjusting the ratio of adversarial examples, we can observe the model's performance under different interference levels in more detail, providing more references for the model's performance evaluation.

4.3.4 Dataset D

In Dataset D, we adjusted the ratio of adversarial examples to normal images to 1:2. 400 normal images and 200 images with adversarial examples were collected for each traffic sign. With this setup, we can further enhance the impact of adversarial examples in the dataset to assess the robustness and adversarial capability of the model more comprehensively.

4.3.5 Dataset E

Finally, in dataset E, we keep the ratio of normal images to adversarial examples at 1:1. 300 normal images and 300 images with adversarial examples are collected for each traffic signpost. This balanced ratio setting is intended to allow the model to have an equal number of training samples in the normal and adversarial cases, thus better evaluating the overall performance and adversarial capabilities of the model.

4.4 Model training

In our research, we conducted model training through the Google Colab platform. We uploaded the organized datasets to Google Drive, which makes data access and management more convenient. We chose PyTorch version 1.11.0 and the corresponding torch vision version 0.12.0 to ensure we could use the functions and optimizations. At the same time, we adopted Python 3.11. In addition, we also use CUDA 12.1 to utilize GPU for accelerated calculations to improve the efficiency of model training.

During the model training process, we divided the Dataset into a training set and a test set to ensure the independence and fairness of training and testing. This method can effectively evaluate the model's generalization ability and reduce the bias introduced by uneven partitioning of the Dataset. We set the number of training epochs to 60, a reasonable value verified through experiments, which can avoid overfitting problems while ensuring model performance. The learning rate is set to 0.001, a commonly used initial learning rate that can maintain a faster convergence speed in the early stages of training and avoid oscillation during the training process. We chose the Adam optimizer, an adaptive learning rate optimization method that can automatically adjust the learning rate to adapt to the characteristics of different parameters, thereby improving the model's convergence speed and generalization ability.

4.5 Evaluation methodology

In terms of evaluation methods, since our dataset is a four-category, the traditional average calculation or simple accuracy evaluation method is not accurate for our statistical experimental results. We chose the four-category confusion matrix method [5, 37]. This evaluation method is very suitable for our experiments. We introduce the four evaluation indicators of precision, recall, F1 Score, and accuracy. The formula is as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} \pm \text{FP}} \tag{1}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} \pm \text{FN}} \tag{2}$$

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

$$\text{accuracy} = \frac{\text{TP}}{\text{Total}} \tag{4}$$

Table 1. The distribution of TP, FN, FP, and TN is forward as the positive classification.

		Predicted			
		Forward	**Left**	**Right**	**Stop**
Actual	Forward	TP		FN	
	Left Right	FP		TN	
	Stop				

In the calculation of evaluation indicators, we introduced four indicators: P (Positive), N (Negative), T (True), and F (False) to evaluate the environmental perception ability of smart cars. Among them, P means that the smart car predicts a positive classification, and N means that the smart car predicts an adverse classification. T means that the smart car's prediction is correct, and F means that the smart car's prediction is wrong. We took the related indicators of the forward class as an example to illustrate. By using forward as a positive classification and other categories as a negative classification, we obtained a 4×4 confusion matrix. In this matrix, TP means that the smart car correctly recognizes the positively classified traffic signs, FP means that the smart car incorrectly recognizes the negatively classified traffic signs as positive classification, TN means that the smart car correctly identifies the negatively classified traffic signs as negative. FN means the smart car mistakenly identifies positively classified traffic signs as negatively classified.

4.6 Experiment

4.6.1 Experiment 1: Evaluation of the effectiveness of different physical attacks in traffic sign attack recognition

This experiment evaluates the effectiveness of using the ResNet18 deep neural network model to test different physical attacks. As a commonly used model for image recognition, ResNet18 has the advantages of depth and residual connection. The model can learn identity mapping and avoid information loss when processing complex image scenes. We fully trained and tuned Dataset A on colab, obtained the deep learning model ResNet18(A), and deployed it to the smart car. Then, we conducted experiments on a series of single-type adversarial attacks, including colored bars, background noise, QR codes, graffiti attacks, PR2, and the mixed attacks we proposed (Figure 10), to compare the performance of ResNet18(A) in different physical the accuracy, recall rate, precision and F1 Score of identifying traffic signs under attacks are used to comprehensively evaluate the performance of the model in the face of various attacks.

Figure 10. Six different physical attacks.

4.6.2 Experiment 2: Comparative analysis of different deep learning models in traffic sign attack recognition

This experiment explores the differences in the performance of different deep learning models for traffic road marking under the same physical attack. In addition to ResNet18, which is mentioned in Experiment 1, AlexNet is a classical model, and CBAM-ResNet26&CBAM-AlexNet is also used. Firstly, we trained and deployed these four models, and then, based on the results of Experiment 1, we applied the most interfering physical attack on traffic road signs to compare the performance of these four models in recognizing the attacked road signs. Performance. By analyzing the performance of the models in depth, we can compare the models that are more resistant to interference.

4.6.3 Experiment 3: Exploring the effectiveness of Channel Attention and Spatial Attention

This experiment explores the effectiveness of Channel Attention and Spatial Attention in CBAM (Convolutional Block Attention Module). To verify whether both need to be applied simultaneously, we designed ablation experiments to remove the Channel Attention or Spatial Attention Module, respectively, and evaluate their effects on the model performance. We chose dataset A, used ResNet26 as the base model, added the following three modules: channel attention, spatial attention, and applied channel and spatial attention at the same time, conducted training and testing, and generated a total of three deep learning models deployed sequentially on a smart vehicle platform for testing and recorded the performance of the models in the mixed attack. The main metrics we focus on include model accuracy and robustness to environmental changes.

4.6.4 Experiment 4: Exploring the optimal ratio of physical adversarial training datasets

This experiment explores the defensive effect of models trained on Datasets with different proportions of adversarial examples against physical attacks. We selected Dataset B, Dataset C, Dataset D, and Dataset E as experimental objects. We used ResNet18, Alex Net, and CBAM-ResNet26 & CBAM-Alexnet to train and test each dataset and generated 12 deep-learning models. They are deployed on smart cars, and then mixed attacks are applied to traffic signs to record the performance of these models for evaluation. During the experiment, we focused on the accuracy of the model on the attacked traffic signs and tested the accuracy of the better-performing model on the original traffic signs. Finally, through the experimental results, we can confirm the good effect of the adversarial training method in physical defense and derive the optimal proportion of adversarial examples in the physical adversarial training set.

5. Result and Discussion

According to Experiment 1, this paper loaded ResNet18(A) trained with Dataset A into the smart car to test the accuracy of applying different physical attacks to traffic signs. The results can be seen in Table 3. The smart car can identify the original traffic signs. The accuracy is 100%, indicating that the model fits well. However, after applying physical attacks, the recognition accuracy decreases significantly. Among them, the mixed attack designed in this paper has the greatest interference, with a recognition accuracy of 38%, which shows that the mixed attack is very effective in attacking smart cars to recognize traffic signs. The background noise has the smallest impact, with an accuracy rate of 98%, indicating that background interference has very little interference with the smart car's recognition of traffic signs, indicating that the smart car's attention is on the pattern of traffic signs. Among single-category physical attacks, the application of QR code attacks also greatly interferes with recognizing traffic signs, with a recognition accuracy of 50%. Followed by color stickers, the recognition accuracy is 75%. Graffiti attack interference is average, and the recognition accuracy is 89%. Then, this paper imitated the robust physical perturbation (RP2) sticker attack proposed by Evtimov et al. and applied it to traffic signs. It was found that the recognition accuracy was not very low, at 67%. This shows that the sticker attack of PR2 is very important to this paper. The experimental interference of multiple traffic signs is not great, and PR2 has certain limitations. The multi-category mixed attacks proposed in this paper will more significantly interfere with smart cars' recognition of traffic signs.

Table 3.b Test results of smart cars against different physical attacks.

Evaluation indicators	Class	Original	Mixed Attack	QR-code	Background noise	Colored bar	Graffiti attack	PR2
Precision	Forward	1	0.44	0.48	0.96	0.72	0.93	0.77
	Left	1	0.42	0.53	1	0.81	0.86	0.68
	Right	1	0.61	0.87	0.96	1	1	0.76
	Stop	1	0.3	0.36	1	0.61	0.79	0.61
Recall	Forward	1	0.33	0.56	1	0.96	0.93	0.8
	Left	1	0.4	0.5	0.96	0.73	0.86	0.66
	Right	1	0.43	0.46	1	0.63	0.86	0.66
	Stop	1	0.46	0.5	96	0.7	0.9	0.7
F1	Forward	1	0.36	0.51	0.97	0.85	0.93	0.78
	Left	1	0.4	0.51	0.97	0.76	0.86	0.66
	Right	1	0.5	0.6	0.97	0.77	0.92	0.7
	Stop	1	0.36	0.41	0.97	0.65	0.84	0.65
Accuracy		**100%**	**38%**	**50%**	**98%**	**75%**	**89%**	**67%**

According to Experiment 2, this paper compares the robustness of three models to physical attacks. We use dataset A to train four models, which are ResNet18 (A), AlexNet (A), CBAM-ResNet26 (A), and CBAM-AlexNet (A), and then from the results of Experiment 1, it can be seen that there is a greater interference with the mixed attack, so this experiment applies the mixed attack is applied to traffic signposts. The experimental results are shown in Table 4. Among the four models, CBAM-ResNet26 (A) shows better anti-interference performance, with a recognition accuracy of 63%, followed by CBAM-AlexNet (A) model, with a recognition accuracy of 51%, which indicates that the introduction of the CBAM module is beneficial to improve the robustness of the model to physical attacks. In contrast, the recognition accuracy of the AlexNet(A) model is 51%. AlexNet(A) model has the lowest recognition accuracy of 30%.

Table 4. Test results of ResNet18(A), Alex Net(A), and CBAM-ResNet26(A)&CBAM-AlexNet(A) on traffic signs applying mixed attacks.

Evaluation indicators	Class	ResNet18(A)	Alex Net(A)	CBAM-ResNet26(A)	CBAM-AlexNet(A)
Precision	Forward	0.44	0.4	0.51	0.52
	Left	0.42	0.3	0.58	0.47
	Right	0.61	0.5	1	0.56
	Stop	0.3	0.18	0.45	0.46
Recall	Forward	0.33	0.4	0.7	0.6
	Left	0.4	0.3	0.56	0.45
	Right	0.43	0.26	0.56	0.47
	Stop	0.46	0.26	0.5	0.5
F1	Forward	0.36	0.4	0.59	0.55
	Left	0.4	0.3	0.56	0.46
	Right	0.5	0.34	0.71	0.51
	Stop	0.36	0.21	0.47	0.48
Accuracy		**38%**	**30%**	**63%**	**51%**

Table 5. Performance comparison of CA-ResNet26(A), SA-ResNet26(A) and CBAM-ResNet26(A).

Evaluation indicators	Class	CA-ResNet26(A)	SA-ResNet26(A)	CBAM-ResNet26(A)
Precision	Forward	0.62	0.6	0.51
	Left	0.58	0.55	0.58
	Right	0.62	0.61	1
	Stop	0.56	0.53	0.45
Recall	Forward	0.65	0.65	0.7
	Left	0.55	0.5	0.56
	Right	0.57	0.55	0.56
	Stop	0.6	0.6	0.5
F1	Forward	0.63	0.62	0.59
	Left	0.56	0.52	0.56
	Right	0.6	0.58	0.71
	Stop	0.58	0.56	0.47
Accuracy		**59%**	**57%**	**63%**

According to Experiment 3, by analyzing the experimental results of the three models, we can confirm that channel attention and spatial attention contribute to the model's performance improvement. According to Table 5, the experiments show that the CBAM model with complete model addition performs best under environmental changes, verifying the necessity of applying channel and spatial attention.

According to experiment 4, we collected dataset B, C, D, and E based on the ratios of 1:5, 1:3, 1:2, and 1:1. Each dataset was trained using ResNet18, Alex Net, and CBAM-ResNet26. A total of 12 deep learning models were generated for testing traffic signs applying mixed attacks. As can be seen from Tables 6, 7, and 8, we found that increasing the proportion of adversarial physical attack samples within a certain range can improve the robustness of the model. The most significant increase in accuracy was observed in the 1:5 and 1:3 ratios. In contrast, the increase in accuracy gradually slowed down in the 1:3 and 1:2 ratios, which demonstrated that physical adversarial training provides a good defense against physical attacks. However, the accuracy of some models decreases compared with the accuracy of the 1:2 ratio and the 1:1 ratio, which makes it difficult to determine the optimal ratio. Therefore, we further analyzed the training effect of the model

at different scales to determine the optimal adversarial physics training scale. We used ResNet18(D) and ResNet18(E) to test the recognition accuracy of original traffic signs. As shown in Table 9, although ResNet18(E) has a slightly higher accuracy than ResNet18(D) for traffic signs with mixed attacks, it does not. According to the original roadmap, ResNet18(D) accuracy is close to 100%. In comparison, the accuracy of ResNet18(E) is 78%, which has declined. When the proportion of attacking samples is too high, it may negatively affect the model's performance, decreasing the model's generalization ability. This suggests that if the proportion of adversarial examples in the total dataset is too high, the model's task of recognizing normal traffic signs will be affected, which confirms that the optimal ratio for physical adversarial training is 1:2 and demonstrates that the adversarial physical training methodology can effectively improve the resilience of the autopilot system against physical attacks.

Table 6. Test results of ResNet18 trained on different proportions of datasets for mixed attacks.

Evaluation indicators	Class	ResNet18(B)	ResNet18(C)	ResNet18(D)	ResNet18(E)
Precision	Forward	0.72	0.74	0.77	0.79
	Left	0.36	0.59	0.78	0.77
	Right	0.68	0.86	0.84	0.83
	Stop	0.3	0.62	0.64	0.69
Recall	Forward	0.52	0.72	0.77	0.75
	Left	0.4	0.65	0.72	0.7
	Right	0.37	0.62	0.7	0.77
	Stop	0.5	0.75	0.8	0.85
F1	Forward	0.6	0.72	0.77	0.76
	Left	0.37	0.61	0.74	0.73
	Right	0.48	0.72	0.76	0.79
	Stop	0.37	0.67	0.71	0.76
Accuracy		**45%**	**68%**	**75%**	**77%**

6. Conclusions

Due to the threat of adversarial attacks on automated driving, this paper is dedicated to studying adversarial physical attacks, physical defense models on automated driving and physical defense training methods. This paper conducts experiments by building an autonomous driving platform. The experimental results show that in terms of adversarial attacks, the mixed attack proposed in this paper performs better than the QR code sticker attack, Colored bar attack, graffiti attack, background attack, and PR2 attack. A mixed attack can reduce the accuracy of traffic sign recognition of an autonomous driving platform by 38%. The CBAM-ResNet26 proposed in this paper has better anti-interference performance regarding physical defense models. Using the same dataset to train and compare the recognition accuracy, the recognition accuracy of CBAM-ResNet26(A) reaches 63%, but the recognition accuracy of Alex Net is only 30%. Furthermore, this paper collects datasets of adversarial physical training with different proportions. Through comparison, it is found that as the proportion of adversarial examples increases, the recognition accuracy of smart cars is higher. However, an excessively high ratio of adversarial examples will reduce the model's recognition accuracy of normal traffic signs. Therefore, this paper finally concluded that the optimal ratio of physical adversarial training is 1:2. The source code and, dataset video for this study can be found at https://github.com/BiChuanxiang/Mixed-Attack-and-Physical-Defense-Strategies.

Table 7. Test results of Alex Net trained on different proportions of datasets for mixed attacks.

Evaluation indicators	Class	Alex Net(B)	Alex Net(C)	Alex Net(D)	Alex Net(E)
Precision	Forward	0.53	0.62	0.66	0.63
	Left	0.34	0.46	0.7	0.64
	Right	0.6	0.84	0.82	0.71
	Stop	0.23	0.5	0.46	0.59
Recall	Forward	0.45	0.57	0.65	0.7
	Left	0.3	0.5	0.6	0.6
	Right	0.35	0.52	0.6	0.62
	Stop	0.4	0.62	0.67	0.65
F1	Forward	0.48	0.59	0.65	0.66
	Left	0.31	0.47	0.64	0.62
	Right	0.44	0.64	0.69	0.66
	Stop	0.29	0.55	0.54	0.61
Accuracy		**37%**	**55%**	**63%**	**64%**
Precision	Forward	0.7	0.81	0.93	0.88
	Left	0.68	0.83	0.86	0.85
	Right	0.87	0.94	1	0.94
	Stop	0.62	0.77	0.79	0.85
Recall	Forward	0.72	0.9	0.93	0.95
	Left	0.65	0.77	0.86	0.87
	Right	0.67	0.8	0.86	0.8
	Stop	0.77	0.87	0.9	0.9
F1	Forward	0.7	0.85	0.93	0.91
	Left	0..66	0.79	0.86	0.85
	Right	0.75	0.86	0.92	0.86
	Stop	0.68	0.81	0.84	0.87
Accuracy		**71%**	**84%**	**89%**	**88%**

Table 9. Test results of CBAM-ResNet26-D and CBAM-ResNet26-E on the original traffic signs.

Evaluation indicators	Class	CBAM-ResNet26(D)	CBAM-ResNet26(E)
Precision	Forward	0.95	0.78
	Left	0.97	0.78
	Right	0.97	0.82
	Stop	0.9	0.73
Recall	Forward	0.95	0.9
	Left	0.92	0.65
	Right	0.95	0.72
	Stop	0.97	0.85
F1	Forward	0.95	0.83
	Left	0.94	0.71
	Right	0.95	0.76
	Stop	0.93	0.78
Accuracy		**95%**	**78%**

7. Acknowledgments

Author Contributions: "Conceptualization, C.X.B., S.S., and J.Q.; methodology, C.X.B., S.S., and J.Q.; software, C.X.B., S.S., and J.Q.; validation, C.X.B., S.S., and J.Q.; formal analysis, C.X.B., S.S., and J.Q.; investigation, C.X.B., S.S., and J.Q.; resources, C.X.B., S.S., and J.Q.; data curation, C.X.B., S.S., and J.Q.; writing—original draft preparation, C.X.B., S.S., and J.Q.; writing—review and editing, C.X.B., S.S., and J.Q.; visualization, C.X.B., S.S., and J.Q.; supervision, J.Q.; project administration, J.Q.; C.X.B contributed 45%, S.S. contributed 5% and J.Q contributed 50%. J.Q is the corresponding author. All authors have read and agreed to the published version of the manuscript."

Funding: C.X.B. and S.S. received full scholarships from CPALL while conducting this research in PIM.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Akhtar N.; Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. *Ieee Access*, **2018**, 6, 14410-14430. https://doi.org/10.1109/ACCESS.2018.2807385

[2] Ni, J.; Chen, Y.; Chen, Y.; Zhu, J.; Ali, D.; Cao, W. A survey on theories and applications for self-driving cars based on deep learning methods.*AS*. **2020**, *10*(8), 2749. https://doi.org/10.3390/app10082749.

[3] Li, Y.; Qu, J. Intelligent road tracking and real-time acceleration-deceleration for autonomous driving using modified convolutional neural networks. *CAST*. **2022**, *22*(6), 10.55003-10.55003 (26 pages). https://doi.org/10.55003/cast.2022.06.22.013.

[4] Bai, T.; Luo, J.; Zhao, J. Inconspicuous adversarial patches for fooling image-recognition systems on mobile devices. *IEEE Internet of Things Journal*. **2021**, *9*(12), 9515-9524. .https://doi.org/10.1109/JIOT.2021.3124815.

[5] Ding, S.; Qu, J. Research on Multi-tasking Smart Cars Based on Autonomous Driving Systems.*SN.* **2023**, *4*(3), 292. https://doi.org/10.1007/s42979-023-01740-1.

[6] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks.*Communications of the ACM*. **2017**, *60*(6), 84-90. https://dl.acm.org/doi/abs/10.1145/3065386

[7] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, **2016**, 770-778.

[8] Li, J.; Chen, X.; Hovy, E.; Jurafsky, D. Visualizing and understanding neural models in NLP. **2015**. arXiv preprint arXiv:1506.01066.https://doi.org/10.48550/arXiv.1506.01066

[9] Cambria, E.; White, B. Jumping NLP curves: A review of natural language processing research.*IEEE Computational intelligence magazine*. **2014**, *9*(2), 48-57.https://ieeexplore.ieee.org/abstract/document/6786458

[10] Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI conference on artificial intelligence*. **2020**, *34*(09), 13693-13696. https://doi.org/10.1609/aaai.v34i09.7123.

[11] Hu, Y. C. T.; Kung, B. H.; Tan, D. S., et al. Naturalistic physical adversarial patch for object detectors. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. **2021**, 7848-7857.

[12] Sharif, M.; Bhagavatula, S.; Bauer, L.; Reiter, M. K. A general framework for adversarial examples with objectives.*ACM Transactions on Privacy and Security (TOPS)*. **2019**, *22*(3), 1-30. https://doi.org/10.1145/3317611.

[13] Xiao, Z.; Gao, X.; Fu, C.; et al. Improving transferability of adversarial patches on face recognition with generative models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. **2021**, 11845-11854.

[14] Kurakin, A.; Goodfellow, I. J.; Bengio, S. Adversarial examples in the physical world. *In Artificial intelligence safety and security*, **2018**.

[15] Evtimov, I.; Eykholt, K.; Fernandes, E.; Kohno, T.; Li, B.; Prakash, A.; Rahmati, A.; Song, D. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*. **2017**, *2*(3), 4.

[16] Athalye, A.; Engstrom, L.; Ilyas, A., et al. Synthesizing robust adversarial examples[C]. *International conference on machine learning. PMLR*, **2018**, 284-293.

[17] Zheng, S.; Song, Y.; Leung, T.; et al. Improving the robustness of deep neural networks via stability training[C]. *Proceedings of the ieee conference on computer vision and pattern recognition*. **2016**, 4480-4488. https://www.cv-foundation.org/openaccess.
[18] Kurakin, A.; Goodfellow, I. J.; Bengio, S. Adversarial examples in the physical world[M]//Artificial intelligence safety and security. *Chapman and Hall/CRC*, **2018**, 99-112.
[19] Madry A. Towards deep learning models resistant to adversarial attacks[J]. *arXiv preprint arXiv:1706.06083*, **2017**.
[20] Wei, X.; Guo, Y.; Yu, J. Adversarial sticker: A stealthy attack method in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **202**2, *45*(3), 2711-2725. https://doi.org/10.1109/TPAMI.2022.3176760.
[21] Wei, X.; Guo, Y.; Yu, J., et al. Simultaneously optimizing perturbations and positions for black-box adversarial patch attacks[J]. *IEEE transactions on pattern analysis and machine intelligence*, **2022**, *45*(7), 9041-9054. https://doi.org/10.1109/TPAMI.2022.3231886.
[22] Wei, X.; Yan, H.; Li, B. Sparse black-box video attack with reinforcement learning. International Journal of Computer Vision. **2022**, *130*(6), 1459-1473. https://link.springer.com/article/10.1007/s11263-022-01604-w.
[23] Goodfellow, I. J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples[J]. *arXiv preprint arXiv:1412.6572*, **2014.** https://doi.org/10.48550/arXiv.1412.6572.
[24] Girshick, R.; Donahue, J.; Darrell, T., et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*. **2014**, 580-587.
[25] Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y. N. Convolutional sequence to sequence learning. *In International conference on machine learning, PMLR;* **2017**.
[26] Graves, A.; Mohamed, A.; Hinton, G. Speech recognition with deep recurrent neural networks. *IEEE international conference on acoustics,* **2013**, 6645-6649. https://doi.org/10.1109/ICASSP.2013.6638947.
[27] McCauley, R. N.; Henrich, J. Susceptibility to the Müller-Lyer illusion, theory-neutral observation, and the diachronic penetrability of the visual input system. *Philosophical Psychology*. **2006**, *19*(1), 79-101. https://doi.org/10.1186/s13229-017-0127-y.
[28] Szegedy C. Intriguing properties of neural networks[J]. *arXiv preprint arXiv:1312.6199*, **2013**.
[29] Le, Q. V. Building high-level features using large scale unsupervised learning. *international conference on acoustics, speech and signal processing. IEEE*, **2013**, 8595-8598. https://doi.org/10.1109/ICASSP.2013.6639343.
[30] Papernot, N.; McDaniel, P.; Goodfellow, I. Jha, S.; Celik, Z. B.; Swami, A. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*. **2016**, *1*(2), 3.
[31] Lu. J.; Sibai, H.; Fabry, E. Adversarial examples that fool detectors[J]. *arXiv preprint arXiv:1712.02494*, **2017**. https://doi.org/10.48550/arXiv.1712.02494.
[32] Duan, R.; Ma, X.; Wang, Y.; Bailey, J.; Qin, A. K.; Yang, Y. Adversarial camouflage: Hiding physical-world attacks with natural styles. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, **2020**.
[33] Lu, J.; Sibai, H.; Fabry, E., et al. No need to worry about adversarial examples in object detection in autonomous vehicles[J]. *arXiv preprint arXiv:1707.03501*, **2017**.
[34] Kurakin, A. Goodfellow I, Bengio S. Adversarial machine learning at scale[J]. *arXiv preprint arXiv:1611.01236*, **2016**.
[35] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses[J]. *arXiv preprint arXiv:1705.07204*, **2017**.
[36] Chen, X.; Li, X.; Zhou, Y., et al. DDDM: a Brain-Inspired Framework for Robust Classification[J]. *arXiv preprint arXiv:2205.10117*, **2022.**
[37] Qu.; J. Multi-Task in Autonomous Driving through RDNet18-CA with LiSHTL-S Loss Function.*ECTI Transactions on Computer and Information Technology (ECTI-CIT)*. **2024**, *18*(2), 158-173. https://doi.org/10.37936/ecti-cit.2024182.254780.