Physical Interference Attacks on Autonomous Driving

Chuanxiang Bi[1] and Jian Qu[2*]
[1,2]Faculty of Engineering and Technology, Panyapiwat Institute of Management,
Nonthaburi, Thailand
E-mail: 6572100057@stu.pim.ac.th, jianqu@pim.ac.th

Received: February 1, 2024 / Revised: April 4, 2024 / Accepted: June 25, 2024

Abstract—**Recent studies have revealed that there are serious security risks to autonomous driving, despite the notable advancements made by deep neural networks in this field. Simple sticker jamming has little experimental validation, despite recent proposals for physical attacks successfully implementing jamming in the real world and misleading autonomous driving recognition. This study focuses on the practicality of various sticker-based physical jammers, such as background noise, colorful stickers, smiley face stickers, and QR code stickers. To boost the study's actual impartiality, we replace the genuine self-driving car in this work with a smart car that performs similar activities. We then utilize three models to train our dataset and carry out five sets of tests. Based on the results, it can be concluded that the QR code sticker has the most potential to interfere with the smart car. This interference causes the smart car's accuracy in recognizing road signs to be between 30% and 40%, whereas the accuracy of the other interferences is over 50%. Furthermore, it demonstrated that, out of the three models, Resnet18 had the best anti-interference capability.**

Index Terms—**Deep Neural Networks, Autonomous Driving, Physical Attacks, Smart Car, Stickers, Resnet18**

I. Introduction

Deep Neural Networks (DNNs) [1] have achieved amazing success in many fields, such as natural language processing [2] and autonomous driving [3]-[7]. However, new research shows that DNNs are vulnerable to adversarial attacks, which can pose significant security risks. Deliberate manipulation of DNN inputs can lead to misbehavior, making adversarial attacks a popular area of academic research with practical implications for real-world applications. In the field of computer vision, adversarial attacks [8]-[10] are now divided into two categories, digital and physical, with the main difference being their different forms. The digital form, where the attacker can feed the input digital image directly into the DNN classifier, also suggests that most digital attacks are white-box attacks, where the attacker needs to know the full details of the model. Digital attacks [11], [12], although they perform well in modeling, are widespread and difficult to identify; they are susceptible to their surroundings, making it difficult to migrate the digital attacks to the physical world. However, physical attacks [13], [14] are carried out in real environments and are therefore more practical and valuable for research and development compared to digital attacks, so more people have gone into physical attacks. But so far, physical attacks on computer vision systems are still very challenging. Physical attacks must be robust enough to withstand variations in illumination, viewing distance, and angle, and image distortion due to camera limitations. There is a limit to the area that can be disturbed by an attacking target. Any background image behind a road sign in a captured image is an example of a disturbance that an algorithm can introduce into a digital image. However, since there is no stable background in the real world, it is not possible to perturb the background there. As a result, only the attacked party itself can be attacked. Furthermore, there are already several available attack techniques; some of them produce complex patterns, while others produce microscopic attacks that are imperceptible to the human eye. A technique for misclassifying printed hostile instances when viewed through a smartphone camera has been demonstrated by Kurakin *et al.* [15]. Alternatively, it is more challenging to apply these techniques in the real world. Others have gone on to attack real stop signs so that self-driving cars do not recognize them correctly and make poor decisions. If the attacker can physically robustly manipulate the road sign, the deep neural network may misclassify it as some other action, which could lead to serious consequences. For example, ShapeShifter [16] uses formula execution to create adversarial stop signs with complex designs, but implementation in real traffic signs is challenging and prone to suspicion.

So, in response to the above problem, this paper focuses on physical attacks that are effective in the physical world, but we are different from most of the research nowadays, which uses physical attacks that are generated in code with targeted attacks, such that the attacks have precision, some of them are so tiny that they may print out with missing pixel dots and lose their effectiveness. Some attacks are large and require the attacker to cover the entire road sign, but such attacks are often too noticeable and cumbersome to implement. The research in this paper is to find physical attacks-physical stickers-which already exist in the real world, and to experiment on traffic road signs with physical stickers through an established autonomous driving platform, and based on the experimental results, to come up with physical sticker attacks that are more threatening to the recognition of the road signs by the self-driving cars.

The contributions of this paper are:

1. This paper proposes that QR code stickers have the strongest ability to interfere with the recognition of road signs by self-driving cars; however, background noise has essentially no influence on the recognition of road signs by self-driving cars.

2. This paper uses the same dataset to train three kinds of deep network models resnet18, mobile net, and Alex net, through the test with physical stickers on the interference of road signs on the self-driving car, the experimental results show that the Resnet18 in the three kinds of models in the strongest anti-jamming ability.

II. Literature Review

A. Adversarial Attack

Adversarial Attacks are purposefully designed input samples that allow machine learning models to misclassify or misjudge. Such attacks may result in a decrease in model performance or may fail. A common application of adversarial attacks is in image classification tasks, where the original image is modified in such a way that the model outputs incorrect classification results by making modifications to the original image that are smart and imperceptible to the human eye. x is the original input. x' in the adversarial sample $x' = x + \delta$, which is obtained by adding a smart perturbation δ. $f(x)$ is the model's output for the original input x. The goal of an adversarial attack is to cause the model to misclassify or miscategorize the original image, but the model's performance may be degraded. The goal of the adversarial attack is to make the output of the adversarial sample x' from the model different from the original input x, so that $\arg\max f(x') \neq \arg\max f(x)$.

Attacks against traffic signs were typically conducted in a white box [17] setting in the early days. Lu et al. [18] attacked the traffic sign detection algorithm. But for the method to work, there had to be significant perturbations because it was not stable enough. Generative Adversarial Networks (GANs) can also generate adversarial instances; however, controlling the generation process of GANs makes it challenging to employ them for focused attacks on certain targets. White-box environments can yield high success rates since they give complete access to the machine-learning model. Nevertheless, the attack method's efficacy sharply declines when it is applied to a black-box model. Black-box [19] assaults are significantly more useful in the real world and have received more practical research than white-box attacks. Black-box attacks, such as generic disturbances, are untargeted assaults that can be employed on any image. The attack strategy presented in this research is also under the category of black-box attacks, which can create an assault that can spoof a target model without the need for previous knowledge of the target model's structure and algorithms.

B. Physical-Realizability of Adversarial Perturbations

The success of adversarial attacks in the real world has been the subject of extensive academic research in recent years. An overview and comparative analysis of recent physical attacks are presented by Wei *et al.* [20]. Physical adversarial samples must be adjusted to varied camera processing and maintain their effectiveness at varying distances, shooting angles, and lighting conditions. Sittawarin *et al.* [21] suggested a technique in a similar study for concealing antagonistic samples on billboards next to traffic signs. Through the categorization attack, they altered the billboard image to make the model's output appear to be a traffic sign. This assault is hard to detect since it tricks not just the machine-learning model but also the human observer. Furthermore, by printing actual-sized road signs on paper and superimposing them over preexisting signs, the RP2 [22] approach can similarly fool DNN classifiers. In this work, we examine practical and successful physical attacks, like those reported in the previous investigations.

III. Methodology

In physical adversarial attacks, to maximize the performance of the model with the stickers on the sample targets and wrong classification results, the cross-entropy loss of the adversarial samples is generally minimized, and the loss function can be generally defined as:

$$J(x')=CrossEntropy(f(x'),y_{target})+\lambda*\mathrm{Re}gularization(\delta)$$

x is the original input, $x' = Stic\mathrm{ker}(x)$ is the adversarial sample, and $f(x)$ is the adversarial sample model output. $CrossEntropy(f(x'), y_{target}$ denotes the cross-entropy loss [14] of the adversarial sample, and y_{target} is the target category set by the attacker. λ is the hyper-parameter used to balance the adversarial loss and the regularization term. $Regularization(\delta)$ is the regularization of the perturbation term, which can be either an L1 or L2 paradigm, to limit the size of the perturbation and prevent over-modification. Minimization of this loss function will cause the model to produce incorrect classification results on adversarial samples, as it takes into account both classification error and a penalty on the size of the perturbation.

The paper focuses on attacking traffic signs, and this use case is chosen because self-driving cars have a larger security problem for adversarial attacks, and the response is obvious: in general, the real STOP road sign is captured by the camera, and then recognized by the DNN, predicted to be a STOP, and performs the STOP action. However, putting a physical sticker on the real STOP road sign will make the DNN recognize it incorrectly, predict it as another road sign, and execute the wrong action, which indicates that the physical sticker poses a threat to the security of autonomous driving. The adversarial physical attack studied in this paper is not to intentionally generate a targeted pattern in the digital world and then print it to cover the original road sign; that way, on the one hand, some tiny attacks will lose pixel points when printed out, which may weaken the effect of the attack. On the other hand, those with a lot of interference would feel unrealistic. However, the physical adversarial attacks studied in this paper adopt those that can appear in the usual world, which are closer to the real world and do not make people suspicious. In this paper, we print and paste some patterns that will not confuse people when they see them, but will be recognized incorrectly by self-driving cars when they see them. As in Fig. 1, QR codes are very common in the real world. Usually, there may be some unqualified individuals placing advertisements on road signs with recognizable QR codes. The purpose is to test whether the DNN identifies the error generated. A smiley face is a sticker that children like. The purpose is to test whether this type of sticker will disrupt the pattern of the road sign and cause the DNN to identify the error. The colored bar is the interference of multiple colors that occur in the real world. The purpose is to test whether DNNs are interfered with by multiple colors, and background noise is the cluttered background that may occur on road signs in the real world, and this purpose explores whether the attention of DNNs is interfered with Overall, this paper aims to explore the physical adversarial attacks that may occur in the real world and to gain further understanding by comparing the effects of these attacks on autonomous driving.

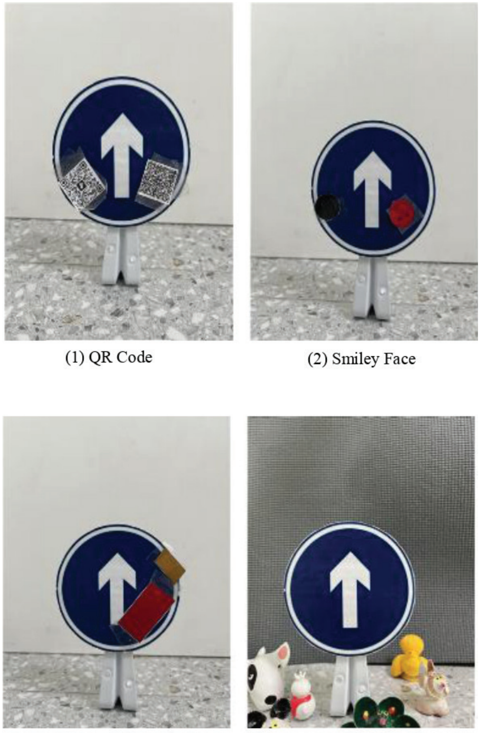

(1) QR Code
(2) Smiley Face
(3) Colored Bar
(4) Background Noise

Fig. 1. Four types of physical stickers

IV. Experimental Setup

A. Experimental Environment

Due to the expensive and safety issues of self-driving cars, this paper selects scale model cars that can be used for self-driving research that simulate self-driving in real scenarios, and uses a Jetson Nano motherboard to make the car an independent agent. The smart car uses a 2,200 mAh battery pack as a power source, and, to approximate the most primitive self-driving car, uses only a camera as an input source to transmit data to the Jetson Nano for processing. The framework of the smart car is shown in Fig. 2. The Jetson Nano is equipped with a driver board that transmits the processing signals from the Jetson Nano to the motor to control the smart car, and incorporates a deep neural network classifier that allows the road sign recognition process to be observed in a more realistic context, thus allowing for a more accurate assessment of the model performance. In terms of the experimental environment, we experiment indoors, and since we want to test whether the physical stickers have an effect on the smart car's recognition of road signs, we have to maintain stable environmental conditions. We used curtains to block the outside light, kept the surroundings unchanged, minimized other noises that could create interference, and kept the indoor lighting stable to create consistent lighting conditions.

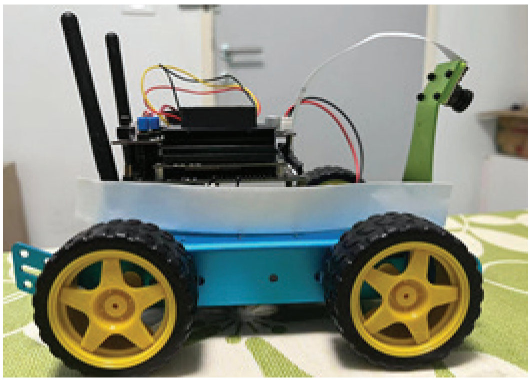
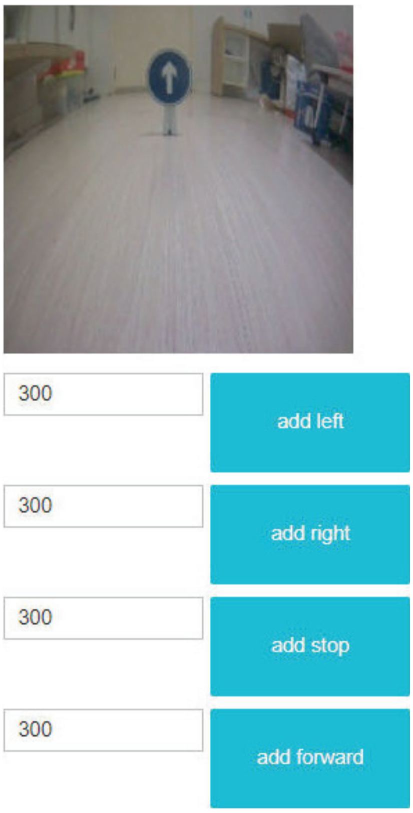


Fig. 2. Frame with Jetson Nano motherboard, and smart car with only one camera as input source

B. Dataset

In order to prevent experimentation by chance, this paper uses 12 road signs, all with the same base and material, to reduce the likelihood that the model will be able to recognize a road sign by observing differences in other aspects of the sign. Diversity is also increased by having different shapes and colors for each category of road signs, making the four categories much less different. With this design, this paper can conduct an effective autonomous driving study in a more realistic environment and improve the robustness of the model for road signs.

Next, in this paper, we use the smart car to collect pictures of the road signs made ourselves, we open the Jupyter and run our code to collect the data, as shown in Fig. 3, as we have four types of road signs about forward, left, right, stop, so the dataset we collect a four-category dataset with forward, left, RIGHT, STOP four categories and the size of the data images is 224*224. We collected 300 images for each of these road signs in the respective category they belong to as shown in Fig. 4. Specifically, we also collected 100 images of no road sign in front in the forward category, to make the cart move forward without road sign in front, in summary, this dataset, there are 3700 images in total, in order to be the model training is better.



300

300

300

300

Fig. 3. Collecting images about four types of road signs: Forward, left, right, and stop, using the Jupyter platform

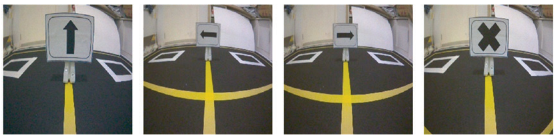


Fig. 4. One of the 4 road signs in the dataset is used as an example

C. Modele Training

In this study, we use Google Colab for training. The collated dataset was uploaded to Google Drive for model training. We adopted the PyTorch framework, specifically using torch 1.11.0, torch vision 0.12.0, Python 3.11, and CUDA 12.1 versions. By training in Colab, we expect to obtain models with good performance.

We also used three deep neural network models for training in our experiments to generate new deep learning models with different performances. These models are resnet18, mobile net, and AlexNet three models, ResNet18 has the advantage of depth and residual connectivity, the model can learn constant mapping, avoiding the loss of information, and maybe more effective in dealing with complex image scenes;

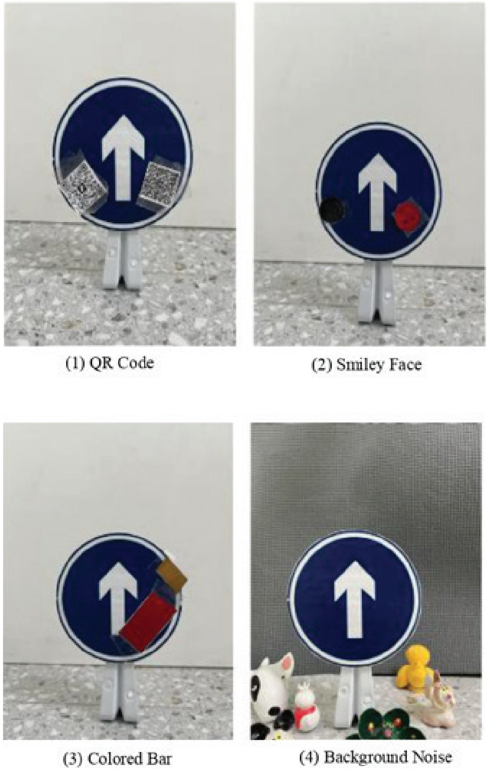
Mobile Net, due to its lightweight design, may be more suitable in resource-constrained environments and So it is more suitable to be used for smart cars in this paper; while AlexNet, as a classical model, has a wide range of applications in tasks such as image classification, target detection, and object recognition, and has better performance in various scenarios, which can be used as a reference benchmark. We chose these three different CNN methods for our experiments, which can evaluate their performance in road sign recognition tasks from different perspectives. Finally, we deploy the trained models to Jetson Nano for model testing and performance evaluation.

D. Experiment

We deployed the three trained models to the Jetson Nano in sequence, after which we tested the following five experiments with a smart car on road signs with different physical stickers, as shown in Fig. 5:

- 12 original road signs without any interference stickers; each road sign is tested ten times.
- 12 road signs with QR code stickers, each road sign is tested ten times.
- 12 road signs with smiley face stickers, each road sign tested ten times.
- 12 road signs with colored stickers; each road sign was tested ten times.
- 12 road signs without any interference stickers, but with ambient noise (background is changing), tested ten times per road sign.

The flow chart is shown in Fig. 6, Each road sign is tested 10 times, each time at a different angle and distance, and As can be seen in Fig. 7, the intelligent model car displays the probability distribution of the actions predicted in real time on our visualization interface when recognizing a road sign while making actions with high probability, and statistically counting the results based on this probability.


(1) QR Code

(2) Smiley Face

(3) Colored Bar

(4) Background Noise

Fig. 5. As an example, one of the forward road signs was sequentially labeled with different physical stickers and placed with background noise

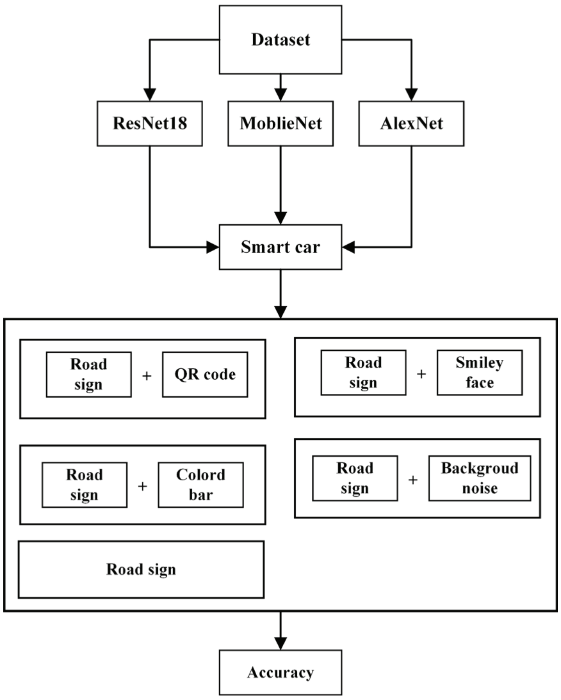
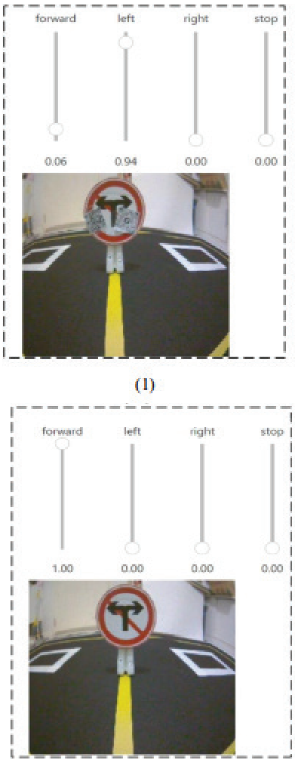


Fig. 6. Experimental framework diagram: the three trained models resnet18, MobileNet, and AlexNet, were deployed to the smart car in turn to test the road sign without interference, the road sign with QR code sticker, the road sign with smiley face sticker, the road sign with colorful sticker, and the road sign with background noise, respectively



(1)

(2)

Fig. 7. The probability distribution of actions displayed on the visualization interface when the intelligent model car recognizes a road sign is plotted: When there is a physical sticker, the forward road sign is predicted to be left with a probability of 0. 94, and when there is no physical sticker, the forward road sign is predicted to be forward with a probability of 1

V. Result

A. Evaluation Methodology

Rather than utilizing the conventional mean calculation method, the experimental evaluation method presented in this study uses a four-category confusion matrix. We have the results from earlier experiments, but there is too much data for a direct comparison. As a result, we decided to combine and condense this fact using a scientific evaluation process. We employed three evaluation measures, which are as follows: precision, recall, F1, and accuracy.

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Reacall} = \frac{TP}{TP+FN}$$

$$F1 = \frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}}$$

$$Accuracy = \frac{Number\ of\ correctly\ classified\ samples}{Total\ number\ of\ samples}$$

Here, we present two metrics—P (Positive) and N (Negative)—for assessing the model vehicle's capacity to perceive its surroundings. First, we designate the remaining classifications as Negative and present one as Positive. Here, P stands for a road sign that is favorably categorized, and N for a road sign that is negatively categorized. The smart car's predictions are shown by the letters T (True) and F (False), respectively, indicating correct and incorrect predictions. For instance, we utilize the other categories as the negative categorization and the LEFT category as the positive categorization to compute the correlation metrics for the LEFT category. When a road sign is correctly classified as positively categorized, the model automobile is said to be in the True Positive (TP) state; when wrongly classed as negatively categorized, it is shown to be in the False Positive (FP) state. The indicators TN (True Negative) and FN (False Negative), respectively, show that the model automobile accurately predicts negatively classified road signs as negatively categorized and incorrectly predicts negatively categorized road signs as favorably categorized. The category to be calculated is the reference for the associated calculation if the pertinent metrics for other categories need to be calculated.

B. Discussions

We performed the statistics by the evaluation methods mentioned above, and as can be seen from Table I, for the original road signs, the accuracy of the smart car recognition is high, as high as 100%, which indicates that the model is well fitted. Then we test several kinds of road signs with physical stickers. The results indicate that the QR code has the greatest interference for the smart car, with the lowest accuracy rate of 40%, and the ambient noise has the least interference for the smart car to recognize the road sign, with the highest accuracy rate of 98%. Smeily Face and Colored bars also have little interference for the smart car to acknowledge the road sign, but they also have a certain degree of interference with an accuracy rate of 89% and 75%.

From Table II and Table III, it can be seen that the same to Table I is that the QR code has the highest interference for the smart car, then Colored bar, Smiley face, and Background noise in that order, where Background noise has the lowest interference, which further indicates that the QR code has the highest interference for the smart car. Moreover, according to the provided experimental results, ResNet18 shows 100% accuracy in the recognition of original road signs, which indicates that ResNet18 has very good fitting ability in the face of simple, undisturbed situations. When confronted with road signs with physical stickers, ResNet18 showed higher accuracy relative to the other models. Even though the QR code caused the most interference for all the models, ResNet18 was able to maintain a relatively high accuracy (40%) in this case, whereas the other models showed a lower accuracy. This leads to the conclusion that ResNet18 is the most resistant to interference, and AlexNet is the least resistant.

TABLE I
EXPERIMENTAL RESULTS OF RESNET18 RECOGNIZING ORIGINAL ROAD SIGNS AND ROAD SIGNS AFFIXED WITH DIFFERENT DISTURBANCES

List	Class	Original	QR	Smiley Face	Colored Bar	Background Noise
Precision	Forward	1	0.44	0.93	0.72	0.96
	Left	1	0.42	0.86	0.81	1
	Right	1	0.61	1	1	0.96
	Stop	1	0.30	0.79	0.61	1
Recall	Forward	1	0.33	0.93	0.96	1
	Left	1	0.40	0.86	0.73	0.96
	Right	1	0.43	0.86	0.63	1
	Stop	1	0.46	0.90	0.70	096
F1	Forward	1	0.36	0.93	0.85	0.97
	Left	1	0.40	0.86	0.76	0.97
	Right	1	0.50	0.92	0.77	0.97
	Stop	1	0.36	0.84	0.65	0.97
Accuracy		**100%**	**40%**	**89%**	**75%**	**98%**

TABLE II
EXPERIMENTAL RESULTS OF MOBLIENET RECOGNIZING ORIGINAL ROAD SIGNS AND ROAD SIGNS AFFIXED WITH DIFFERENT DISTURBANCES

List	Class	Original	QR	Smiley Face	Colored Bar	Background Noise
Precision	Forward	0.85	0.33	0.74	0.51	0.82
	Left	0.87	0.36	0.73	0.58	0.81
	Right	0.89	0.54	0.80	1	0.89
	Stop	0.85	0.23	0.65	0.45	0.84
Recall	Forward	0.96	0.30	0.86	0.70	0.93
	Left	0.90	0.30	0.73	0.56	0.86
	Right	0.83	0.40	0.70	0.56	0.83
	Stop	0.76	0.36	0.63	0.50	0.73
F1	Forward	0.90	0.31	0.79	0.59	0.87
	Left	0.88	0.32	0.73	0.56	0.83
	Right	0.85	0.45	0.74	0.71	0.85
	Stop	0.80	0.28	0.63	0.47	0.78
Accuracy		**86%**	**34%**	**73%**	**58%**	**84%**

TABLE III
EXPERIMENTAL RESULTS OF ALEXNET RECOGNIZING ORIGINAL ROAD SIGNS AND ROAD SIGNS AFFIXED WITH DIFFERENT DISTURBANCES

List	Class	Original	QR	Smiley Face	Colored Bar	Background Noise
Precision	Forward	0.83	0.40	0.77	0.48	0.81
	Left	0.82	0.30	0.68	0.53	0.76
	Right	0.96	0.50	0.76	0.87	0.96
	Stop	0.78	0.18	0.61	0.36	0.78
Recall	Forward	0.86	0.40	0.80	0.56	0.86
	Left	0.80	0.30	0.66	0.50	0.76
	Right	0.86	0.26	0.66	0.46	0.83
	Stop	0.86	0.26	0.70	0.50	0.83
F1	Forward	0.84	0.40	0.78	0.51	0.83
	Left	0.80	0.30	0.66	0.51	0.76
	Right	0.90	0.34	0.70	0.60	0.89
	Stop	0.81	0.21	0.65	0.41	0.80
Accuracy		**85%**	**30%**	**70%**	**50%**	**82%**

VI. Conclusion

With the development of automated driving technology, road sign recognition as one of the key tasks in the automated driving system, so there are many researches to create interference for road sign recognition, but there are some digital attacks that cannot be migrated to the real world, which leads to failure, and there are also some physical attacks that create a larger interference, which is easy to be detected. We have done this by finding out that there are various physical interferences in the real world, such as QR codes affixed and graffiti painted, which may negatively affect the accuracy of road sign recognition. In this paper, we test the interference of QR code stickers, smiley face stickers, colored stickers, and background noise on the recognition of road signs by using a smart car, and the experimental results show that the physical interference of QR code stickers has the most significant impact on the accuracy of road sign recognition by the smart car, and the recognition accuracy is only 30% to 40%, which is much lower than that of the smart car. 30% to 40%, much lower than other types of interference. This suggests that QR code interference may cause serious safety hazards to the autonomous driving system and requires special attention and targeted solutions. In addition, the impact of other types of physical interference on road sign recognition is relatively small, and the recognition accuracy is kept above 50%, but there is still a certain degree of influence. And, in the in-depth comparison of the performance of the three different models, we further confirm the excellent performance of the ResNet18 model in resisting interference. This indicates that ResNet18 performs well in coping with physical interference in the street sign recognition task in the face of different types of physical interference.

In summary, the results of this study highlight the important impact of physical interference on the performance of road sign recognition in automated driving systems, especially the severity of QR code interference. Future research can further explore the mechanism of different types of physical interference on road sign recognition, as well as the future development towards physical defense that can be made to cope with the impact of physical interference on the road sign recognition performance of an automated driving system. All these will help to improve the robustness and safety of the system, and promote the development of automatic driving technology towards a more mature and reliable.

Acknowledgment

Conceptualization, CX. B. and J. Q.; methodology, CX. B. and J. Q.; software, CX. B. and J. Q.; validation, CX. B. and J. Q.; formal analysis, CX. B. and J.Q.; investigation, CX. B. and J. Q.; data curation, CX. B. and J. Q.; writing—original draft preparation, CX. B. and J. Q.; writing—review and editing, CX. B. and J. Q.; visualization, CX. B. and J. Q.; supervision, J. Q.; All authors have read and agreed to the published version of the manuscript. J. Q. is the corresponding author.

The first author received scholarship support from CPALL for conducting this research in PIM.

References

[1] A. Aldahdooh, W. Hamidouche, S. A. Fezza, and O. Deforges, "Adversarial example detection for DNN models: A review and experimental comparison," *Artif Intell Rev*., vol. 55, no. 6, pp. 4403-4462, May 2022.

[2] J. Li, X. Chen, E. Hovy, and D. Jurafsky, "Visualizing and understanding neural models in NLP," *arXiv*, 2015. [Online]. Available: https://arxiv.org/abs/1506.01066 [Accessed: Jan. 14, 2024].

[3] J. Qu and S. Shi, "Multi-Task in autonomous driving through RDNet18-CA with LiSHTL-S loss function," *ECTI-CIT*, vol. 18, no. 2, pp. 158-173, Apr. 2024.

[4] Y. Li and J. Qu, "A novel neural network architecture and cross-model transfer learning for multi-task autonomous driving," *Data Technologies and Applications*, vol. 58, no. 5, pp. 693-717, Jan. 2024, https://doi.org/10.1108/DTA-08-2022-0307

[5] S. Ding and J. Qu, "Automatic driving for road tracking and traffic sign recognition," *STA*, vol. 27, no. 4, pp. 343-362, Dec. 2022.

[6] Y. Li and J. Qu, "Intelligent road tracking and real-time acceleration-deceleration for autonomous driving using modified convolutional neural networks," *Curr. Appl. Sci. Technol.*, vol. 22, no. 6, pp. 1-26, Mar. 2022..

[7] S. Ding and J. Qu, "Research on multi-tasking smart cars based on autonomous driving systems," *SN Computer Science*, vol. 4, no. 3, p. 292, Mar. 2023.

[8] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1712.09665 [Accessed: Jan. 14, 2024].

[9] Y. Dong, F. Liao, T. Pang et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9185-9193.

[10] A. Madry, A. Makelov, L. Schmidt et al., "Towards deep learning models resistant to adversarial attacks," *arXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1706.06083 [Accessed: Jan. 14, 2024].

[11] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Priv. (SP)*, 2017, pp. 39-57.

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

[13] R. Duan, X. Ma, Y. Wang et al., "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1000-1008.

[14] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410-14430, Jan. 2018.

[15] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Artif. Intell. Safety Security*, 2018, pp. 99-112.

[16] S. T. Chen, C. Cornelius, J. Martin, and D. Horng Chau, "Shapeshifter: Robust physical adversarial attack on Faster R-CNN object detector," in *Proc. Mach. Learn. Knowl. Discovery Databases: European Conf.*, 2019, pp. 52-68.

[17] S. Chow, P. Eisen, H. Johnson, and P. C. van Oorschot, "White-box cryptography and an AES implementation," in *Proc. 9th Annu. Workshop Sel. Areas Cryptogr. (SAC)*, 2002, pp. 250-270.

[18] J. Lu, H. Sibai, and E. Fabry, "Adversarial examples that fool detectors," *arXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1712.02494 [Accessed: Apr. 14, 2024].

[19] S. Nidhra and J. Dondeti, "Black box and white box testing techniques-A literature review," *Int. J. Eng. Sci. Appl.*, vol. 2, no. 2, pp. 29-50, Jun. 2012.

[20] W. Hui, "Physical adversarial attack meets computer vision: A decade survey," *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2209.15179 [Accessed: Jan. 20, 2024].

[21] C. Sitawarin, A. N. Bhagoji, A. Mosenia, and P. Mettal, "Rogue signs: Deceiving traffic sign recognition with malicious ads and logos," *arXiv*, 2018. [Online]. Available: https://arxiv.org/abs/1801.02780 [Accessed: Apr. 15, 2024].

[22] I. Evtimov, K. Eykholt, E. Fernandes et al., "Robust physical-world attacks on machine learning models," *arXiv* , 2017. [Online]. Available: https://arxiv.org/abs/1707.08945 [Accessed: Jan. 20, 2024].

[23] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *arXiv*, 2018. [Online]. Available: https://arxiv.org/abs/1805.07836 [Accessed: Jan. 20, 2024].

Chuanxiang Bi is currently studying for the Master of Engineering Technology, Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand. He received B.B.A from Nanjing Tech University Pujiang Institute, China, in 2022. His research interests are Research direction is artificial intelligence, image processing, and autonomous driving.

Jian Qu is an Assistant Professor at the Faculty of Engineering and Technology, Panyapiwat Institute of Management. He received a Ph.D. with an Outstanding Performance award from Japan Advanced Institute of Science and Technology, in 2013. He received a B.B.A with Summa Cum Laude honors from the Institute of International Studies of Ramkhamhaeng University, Thailand, in 2006, and M.S.I.T from Sirindhorn International Institute of Technology, Thammasat University, Thailand, in 2010. He has been serving on a house committee for the Thai Superai project since 2020. His research interests are natural language processing, intelligent algorithms, machine learning, machine translation, information retrieval, image processing, and autonomous driving.