Research on Adversarial Patch Attack Defense Method for Traffic Sign Detection

Yanjing Zhang[1], Jianming Cui[1], and Ming Liu[2](✉)

[1] School of Information Engineering, Chang'an University, Shaanxi 710064, China
[2] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China
liuming@cert.org.cn

Abstract. Accurate and stable traffic sign detection is a key technology to achieve L3 driving automation, and its performance has been significantly improved by the development of deep learning technology in recent years. However, the current traffic sign detection has inadequate difficulty resisting anti-attack ability and even does not have basic defense capability. To solve this critical issue, an adversarial patch attack defense model IYOLO-TS is proposed in this paper. The main innovation is to simulate the conditions of traffic signs being partially damaged, obscured or maliciously modified in real world by training the attack patches, and then add the attacked classes in the last layer of the YOLOv2 which are corresponding to the original detection categories, and finally the attack patch obtained from the training is used to complete the adversarial training of the detection model. The attack patch is obtained by first using RP_2 algorithm to attack the detection model and then training on the blank patch. In order to verify the defense effective of the proposed IYOLO-TS model, we constructed a patch dataset LISA-Mask containing 50 different mask generation patches of 33000 sheets, and then training dataset by combining LISA and LISA-Mask datasets. The experiment results show that the mAP of the proposed IYOLO-TS is up to 98.12%. Compared with YOLOv2, it improved the defense ability against patch attacks and has the real-time detection ability. It can be considered that the proposed method has strong practicality and achieves a tradeoff between design complexity and efficiency.

Keywords: Traffic sign detection · Adversarial patch attack · Deep learning

1 Introduction

Traffic sign detection is a key technology that is continuously updated and iterated in the vision-based advanced driver assistance systems. Its purpose is to establish accurate, real-time and safe traffic sign recognition capabilities for complex and dynamic real roads [1]. The most widely used technology is target detection based on Deep Neural Networks (DNN) [2]. However, many recent studies have shown that the security of DNN models is not reliable, that is, it is susceptible to the influence of adversarial examples, which would mislead the classifier produces incorrect predictive output [3–5]. Currently, adversarial patch attacks in the physical world have been considered as a very effective

W. Lu et al. (Eds.): CNCERT 2022, CCIS 1699, pp. 199–210, 2022.
https://doi.org/10.1007/978-981-19-8285-9_15

means for attacking object detection models, and have achieved remarkable results in the fields of image classification [6], face recognition [7], object detection and etc. [8–10]. In order to deal with the security threats caused by patch attacks, a growing number of researchers began to study defense methods. However, current researches mainly focus on image classification, and there are few reports on traffic sign detection. In addition, traditional image pre-processing methods, such as image denoising [11], local gradient smoothing [12], and partial occlusion [13], would reduce the detection accuracy on the original samples, and most of them are designed to operate in the digital space and are ineffective to the physical world.

YOLO (You Only Look Once) series is a one-stage object detector that can directly output bounding boxes and categories. Compared with RCNN (Region-Convolutional Neural Networks), Faster-RCNN and other two-stage networks, YOLO has a lighter structure, fewer parameters, and faster speed. Therefore, it is more suitable for application research in the field of automatic driving that requires high real-time and accuracy [14]. Compared with v3–v5, YOLOv2 has less computation in forward reasoning [15–18], and can maintain a relatively high mAP (mean Average Precision) in the COCO dataset test under the same scale input. In addition, in automatic driving, object detection models are mostly deployed on edge devices for inference, resulting in limited model storage space and computing resource [19]. YOLOv2 mainly consists of convolutional layers and softmax, which is easier to implement in mobile device and can also accelerate inference by small graphics cards. Therefore, the interesting and challenging question addressed here is how to integrate and extend YOLOv2 to traffic sign detection and achieved the stable defense capability.

To solve the above problems, we propose an adversarial patch defense model IYOLO-TS (Improved YOLOv2 on Traffic Signs) on traffic sign detection. The main contributions can be summarized as follows: (1) We extend the research of patch attack defense to the field of traffic sign detection and proposed a practical defense model IYOLO-TS. (2) We improved the last layer of YOLOv2 model by adding an additional 11 attacked classes, and optimized it structure to ensure the high detection performance for normal traffic signs. (3) In order to achieve high robustness and more realistic style against perturbations, we adopt RP_2 algorithm [8] to attack the YOLOv2 and pioneered the development of a patch dataset named LISA-Mask.

2 Improved YOLOv2 on Traffic Signs Detection Model

2.1 Framework Design of IYOLO-TS

Figure 1 provides an overview of IYOLO-TS. From the structure of the neural network, IYOLO-TS adds 11 additional attacked categories to the last softmax layer. As a result, IYOLO-TS is able to detect the attacked targets while accurately identify the attacked targets to the true classes, which are defined as the right part of Fig. 1.

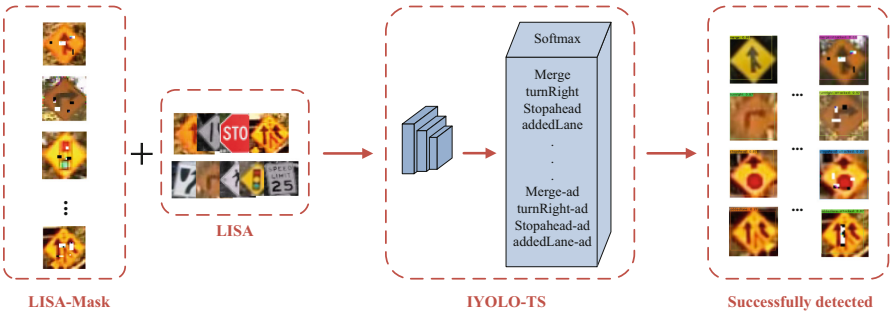

Fig. 1. Framework of IYOLO-TS.

We sample from each category of LISA and LISA-mask to train IYOLO-TS. IYOLO-TS retains the network structure of yolov2 except for the final softmax layer by adding 11 attacked categories. The right part of figure is attacked traffic sign detection result of IYOLO-TS. The base idea of YOLOv2 is to represent the output of the feature map as the center, width and height of the bounding box, as well as the confidence and category. YOLOv2 divides the input image x into N preselected areas, and each area predicts M anchor box. Assuming that there are n classes to be identified, for the LISA and LISA-Mask datasets, n is 11, each anchor box can be written as an $(n+5)$ dimensional vector. The result of the feature map for each anchor box can be expressed as shown below:

$$\left\langle \hat{X}, \hat{Y}, W, H, P_{obj}, P_{cls1}, \ldots, P_{clsn} \right\rangle \tag{1}$$

where $\hat{X}$, $\hat{Y}$, W, H are the center and size of bounding box, P_{obj} is the confidence score indicates the probability of whether the bounding box contains a target and $P_{cls\mathrm{i}}$ is class score. Then, arrange the anchor boxes in order, and each preselected area would output a vector with dimension $M(n+5)$. Eventually, the output of YOLOv2 is a vector of dimension $NM(n+5)$. IYOLO-TS inherits the form of the YOLOv2 loss function and adds the loss to the attacked class score. We add 11 attacked categories to the last softmax layer of YOLOv2, so the length of each anchor boxes vector becomes $(n+5+11)$, and the corresponding final output becomes a vector of $NM(n+5+11)$ dimension. This gives IYOLO-TS two advantages: the detection speed inherited from YOLOv2 meets the time-sensitive requirements for defending against physical world attacks and can also be used as a model for detecting attacks.

2.2 RP_2-Based Attacking Process

In order to achieve a high robustness and a more realistic style against perturbations, we use the method in [8] to attack the YOLOv2 detectors. To generate visual adversarial perturbations that are robust under different physical conditions, RP_2 algorithm is first derived without considering other physical conditions, starting with the optimal method for generating perturbations to a single image x. Then update the algorithm considering continuous changes in the distance and angle of the camera to the road sign. Then, the

constrained optimization problem of RP_2 is expressed as below:

$$arg \ \min_{\delta} \lambda \|\delta\|_p + J\left[f_\theta(x+\delta), y^*\right] \tag{2}$$

where $J(\cdot)$ is the loss function measures the degree of difference between the prediction of the model and the target class y^*. x is the input, δ denotes the perturbation of input x, $f_\theta(\cdot)$ denotes the target classifier, and λ is the hyperparameter that controls the regularization of the distortion. Specifying the distance function as $\|\delta\|_p$, which denotes the p-norm of δ. To better capture the effects of changing physical conditions, partial experimental samples containing random noise are generated to be added to the algorithm iterations. To ensure that the perturbation is applied only to the surface of the target object, a mask is introduced that will limit the physical region of the perturbation. The final robust spatially constrained perturbation is optimized as:

$$arg \ \min_{\delta} \lambda \|M_x \cdot \delta\|_p + NPS + E_{x_i \sim X^v} J\left\{f_\theta[x_i + T(M_x \cdot \delta)], y^*\right\} \tag{3}$$

where the matrix M_x is the representation of the mask, NPS is the unprintability fraction, and the function $T(\cdot)$ represents the alignment function that maps the transformation of the object and the perturbation. Since all perturbation values must be reproducible in the physical world and there exist some reproduction errors in the colors produced by the printer [20], RP_2 adds an additional term NPS to the objective function to model the printer color reproduction errors. It can be found that during an attack, forged patches generated under the qualification of different masks can simulate common vandalism behaviors that are ignored by most people. Such attacks in the physical world are highly disruptive to traffic sign detectors, so it is imperative to develop appropriate defense strategies.

2.3 Generating of LISA-Mask Dataset

In order to make IYOLO-TS more generalizable and make it effective in defending against various patch attacks, we generate 50 different masks and constructs a new dataset named LISA-Mask to help train the IYOLO-TS.

During attack patches generating experiment, we found that the patches at different locations have an impact on the effectiveness of the attack, and each mask produces a different attack effect. In addition, in order to simulate a more realistic random attack scenario as much as possible, 50 different masks are produced in this paper by limiting the size, distance, number and shape of the scope. The generated masks are different from other target detection datasets that can take the whole area as the area of interest for the attack, the masks in this paper should limit the size of the scope so that they avoid obscuring the whole pattern of traffic signs.

The success rate of the attack can be expressed as follows:

$$\frac{\sum_{c \in C}\left\{f_\theta\left[A\left(c^{d,g}\right)\right] = y^* \wedge f_\theta\left(c^{d,g}\right) = y\right\}}{\sum_{c \in C}\left[f_\theta\left(c^{d,g}\right) = y\right]} \tag{4}$$

where $A(c^*)$ represent a set of images with incorrect classification results from original images set c. $c^{d,g}$ represent the images taken from distance d and angle g. Respectively,

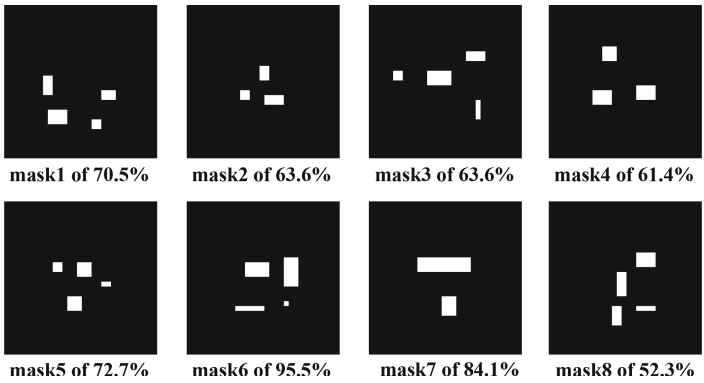

Fig. 2. Some of the masks and their attack success rates.

y is the actual class label of the target, and y^* is the detection result of the target after the attack. As shown in Fig. 2, some of the generated masks and their attack success rates. It can be seen that different kinds of masks can lead to different degrees of reduction in YOLO's inference results, i.e., physical attacks on traffic signs can be simulated to some extent.

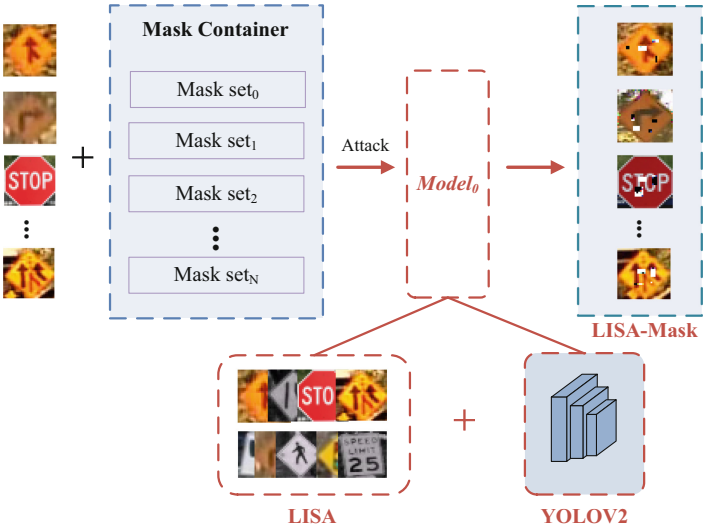

Fig. 3. The generation process of the LISA-Mask dataset.

Figure 3 exhibited the generation process of LISA-Mask dataset. First, YOLOv2 is trained on LISA training set and named as $Model_0$, then 50 different masks are generated by using the aforementioned method, and then the attack on $Model_0$ is performed on different masks based on the method in [8], respectively, the difference of the detection results with the true labels is added to the loss function, and the attack patches are

updated by back-propagation training. The generated patches are applied on LISA, and the images with the patch attacks are obtained, that is named as LISA-Mask dataset. The produced dataset contains a total of 11 categories of traffic sign images, each contained 3000 images that were attacked 50 times, for a total of 33,000 images.

3 Experiments and Results

3.1 Test Bench Setup

To evaluate our proposed work, we constructed the experimental data according to the structure in Fig. 4. Firstly, the LISA-Mask and LISA data sets are merged. There are 11 types of targets and each type of target is divided into clean data and attacked data. Then, to keep data balance in training, three enhancement methods is used on categories less than 100 pictures in the LISA dataset: contrast, brightness and sharpness change. We don't recommend using cutting, mirroring, rotation and other enhancement methods, for these complex situations are not common in driving detection task. Finally, we selected two hundred images randomly from each category of data to construct the experimental dataset, which is split into 80% training and 20% test set.

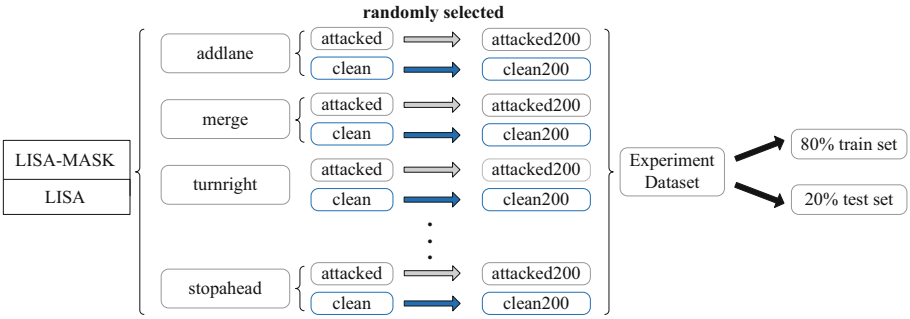

Fig. 4. Construction structure of the experimental dataset.

For all experiment, we use tensorflow1.14 and P4000 for training. YOLO is trained by Adam optimizer with learning rate 0.01, and batch size is 32. In the training of adversarial patches, SGD is used with learning rate 0.01, and decay rate is set to 0.1.

3.2 Object Detection

Object Selection Performance Analysis of IYOLO-TS on Clean Dataset
To evaluate the performance of IYOLO-TS, we calculate the AP of YOLOv2 and IYOLO-TS for each class on the LISA test set in Table 1. It can be observed that IYOLO-TS has less reduced in AP for each class compared to YOLOv2. On average, the mAP of IYOLO-TS is 97.75%, which is only 1.25% lower compared to YOLOv2, indicating that IYOLO-TS can maintain a strong roadmap detection.

Table 1. Performance of YOLOv2 and IYOLO-TS on the LISA test set

Classes	addlane	keepright	laneend	merge	signalahead	limit25
YOlOv2	100%	100%	92.39%	98.53%	100%	100%
IYOLO-TS	100%	100%	91.76%	96.49%	100%	98.63%
	limit30	limit45	stopahead	stop	turnright	mAP
	100%	100%	100%	100%	100%	99.00%
	100%	100%	100%	100%	100%	97.75%

Analysis of the Validity of IYOLO-TS Defense Detection

To evaluate the defensive capability of IYOLO-TS, we calculated AP of each class on the dataset. It can be seen that IYOLO-TS can distinguish the adversarial samples from the clean data, and the mAP reaches 98.12%. Table 2 shows the detection AP of IYOLO-TS for all classes of images, and it can be seen that IYOLO-TS has a strong defense detection performance. Figure 5 shows the performance of IYOLO-TS and YOLOv2 against patch attacks. As can be seen that, compared to YOLOv2, IYOLO-TS achieves higher metrics in all the other 10 classes of flags except the signalahead class, which shows a stronger defense against attacked data.

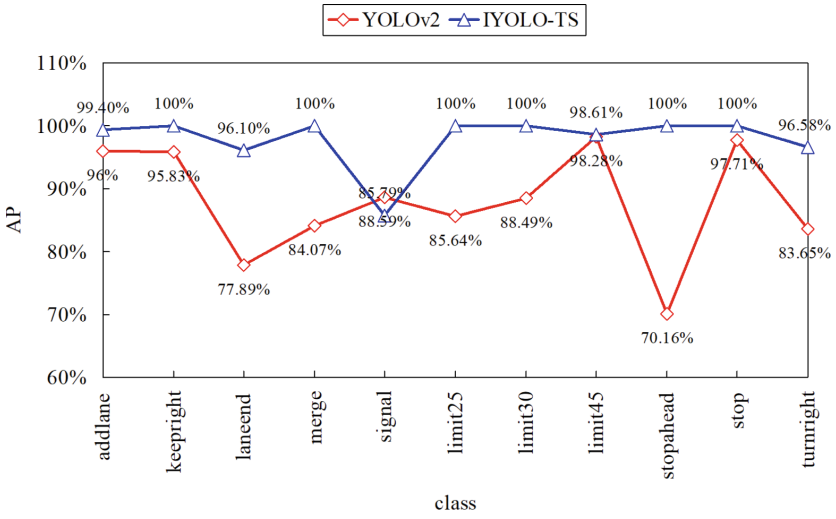

Fig. 5. Performance comparison of IYOLO-TS and YOLOv2 against patch attacks.

Figure 6 shows the defense effect on LISA-Mask. The attacked addedlane is able to successfully trick YOLOv2 to identify it as the merge class, however, IYOLO-TS is able to successfully and correctly identify the attacked target.

Table 2. IYOLO-TS AP for 22 classes of images, where classes indicate clean traffic signs data and classes-ad indicate attacked traffic signs data

Classes	AP	Classes-ad	AP
addlane	96.34%	addlane-ad	100%
keepright	100%	keepright-ad	100%
laneend	100%	laneend-ad	100%
merge	91.46%	merge-ad	97.67%
signalahead	92.5%	signalahead-ad	93.51%
limit25	100%	limit25-ad	100%
limit30	100%	limit30-ad	100%
limit45	95.74%	limit45-ad	95.65%
stopahead	100%	stopahead-ad	100%
stop	100%	stop-ad	95.45%
turnright	97.37%	turnright-ad	100%
mAP	98.12%		

Fig. 6. Performance of YOLOv2 and IYOLO-TS for detection of attacked added lane.

In addition, IYOLO-TS adds 11 additional attacked classes to the structure of YOLOv2, as Fig. 7 shows the detection results of some of the attacked classes. It can be seen that IYOLO-TS is not only able to correctly identify the attacked traffic sign, but also distinguish whether the traffic sign is under attack or not. It shows that IYOLO-TS has good detection ability for different kinds of patch attacks.

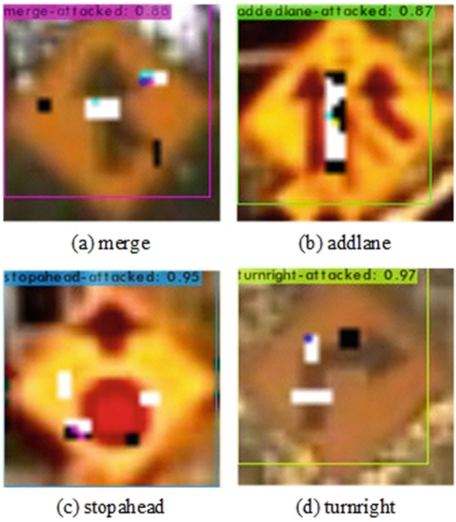

Fig. 7. Detection results of partially attacked classes.

3.3 Analysis of the Effectiveness of Patch Attack Defense

In order to evaluate the defensive capability of IYOLO-TS, we test IYOLO-TS under white-box attacks and physical world attacks respectively.

Defense Effectiveness Analysis under White-box Attacks
We continue with the LISA-Mask generation process, by using RP_2 to generate the patch dataset LISA-$Mask_0$ against IYOLO-TS. First, IYOLO-TS was trained on the LISA training set, and then images with the patch attack were generated on the LISA dataset using RP_2 against the trained IYOLO-TS to obtain the LISA-$Mask_0$ dataset. Then, the generated patch dataset LISA-$Mask_0$ was used to test the IYOLO-TS model. Table 3 shows the performance of IYOLO-TS against white-box attacks.

Table 3. Detection effectiveness of IYOLO-TS against white-box attacks

Classes AP	addlane 95.95%	keepright 100%	laneend 90.79%	merge 90.24%	signalahead 100%	limit25 100%
	limit30 94.37%	limit45 95.35%	Stopahead 98.61%	stop 100%	turnright 100%	mAP 97.22%

As can be seen from the Table 3, except for laneend and merge, which have an accuracy of about 90%, other classes have AP values higher than 94%, indicating that IYOLO-TS still shows a strong defense capability in the face of new attacks.

Defense Effectiveness Analysis under Physical World Attacks
To verify the usefulness of the model in this paper, the defensive performance of IYOLO-TS in the physical world was tested. In the experiments, the generated adversarial patches are printed and attached to the traffic signs to further compare and demonstrate the defense effectiveness of YOLOv2 and IYOLO-TS. As shown in (a) (d) (g) (j) of Fig. 8, YOLOv2 miscalculates under the generated adversarial patch, and the performance of (b) (c) (e) (f) (h) (i) (k) (l) shows that IYOLO-TS can distinguish the clean data from the attack data under physical attacks.

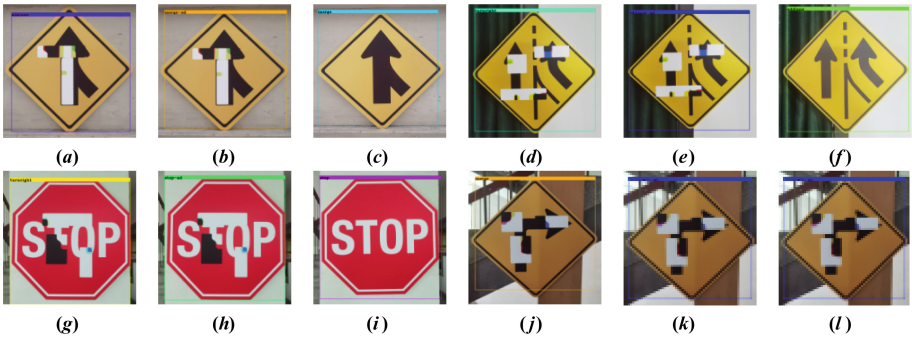

Fig. 8. Physical world attack test sample.

4 Conclusion and Future Work

In this paper, an improved defense model, IYOLO-TS, was firstly proposed to improve the anti-attack ability of the traffic sign detection. Firstly, the masks under multi-scale and multi-constraint conditions were built to simulate random multi-type physical attacks in the physical world, and the first test data set, Lisa-Mask is constructed through annotation fusion. On this basis, 11 attacked classes are innovatively added to the YOLOv2 network structure, so that the model can distinguish the attack samples from the original samples while maintaining the detection capability. In the experiment, we compared the detection performance of IYOLO-TS and YOLOv2, and completed the performance test and analysis of white-box attack and physical world attack respectively. Experimental results show that IYOLO-TS has a good defense ability against the adversarial patch attack from the physical world. But it can also be found that the real road traffic signs obscured, to be damaged, is far beyond this study at this stage can simulate. In addition, vehicle speed, weather, light and other factors will directly affect the processing efficiency of the model. Therefore, in our next work, how to optimize the model to adapt dynamic environment and achieve a more accurate and interpretable detection method are also important and interesting research topics.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (Grant No. 62106060).

References

1. Balasubramaniam, A., Pasricha, S.: Object Detection in Autonomous Vehicles: Status and Open Challenges. arXiv preprint arXiv:2201.07706 (2022)
2. Salah Zaki, P., Magdy William, M., Karam Soliman, B., Gamal Alexsan, K., Khalil, K., El-Moursy, M.: Traffic Signs Detection and Recognition System using Deep Learning. arXiv preprint arXiv:2003.03256 (2020)
3. Yi Huang, Wai-Kin Kong A.: Transferable Adversarial Attack based on Integrated Gradients. arXiv preprint arXiv:2205.13152 (2022)
4. Cilloni, T., Walter, C., Fleming, C.: Focused Adversarial Attacks. arXiv preprint arXiv:2205.09624 (2022)
5. Mo, Z., Patel, V.M.: On Trace of PGD-Like Adversarial Attacks. arXiv preprint arXiv:2205.09586 (2022)
6. Subramanya, A., Pillai, V., Pirsiavash, H.: Fooling Network Interpretation in Image Classification. arXiv preprint arXiv:1812.02843 (2019)
7. Singh, I., Araki, T., Kakizaki, R.K.: Powerful Physical Adversarial Examples Against Practical Face Recognition Systems. arXiv preprint arXiv:2203.15498 (2022)
8. Eykholt, K., et al.: Robust physical-world attacks on deep learning visual classification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1625–1634 (2018)
9. Thys, S., Ranst, W., Goedeme, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 49–55 (2019)
10. Lee, M., Zico Kolter, J.: On physical adversarial patches for object detection. arXiv preprint arXiv:1906.11897 (2019)
11. Hayes, J.: On visible adversarial perturbations & digital watermarking. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Work-shops (CVPRW), pp. 1597–1604 (2018)
12. Naseer, M., Khan, S., Porikli, F.: Local gradients smoothing: defense against localized adversarial attacks. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1300–1307 (2019)
13. McCoyd, M., et al.: Minority reports defense: defending against adversarial patches. arXiv preprint arXiv:2004.13799 (2020)
14. Wang, J., Chen, Y., Gao, M., Dong, Z.: Improved YOLOv5 network for real-time multi-scale traffic sign detection. arXiv preprint arXiv:2112.08782 (2021)
15. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. CoRR (2016)
16. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
17. Bochkovskiy, A., Wang, C.Y., Liao, H.: YOLOv4: Optimal Speed and Accuracy of Object Detection (2020)
18. Ge, Z., Liu, S., Wang, F., et al.: YOLOX: Exceeding YOLO Series in 2021 (2021)
19. Levering, A., Tomko, M., Tuia, D., Khoshelham, K.: Detecting Unsigned Physical Road Incidents from Driver-View Images. arXiv preprint arXiv:2004.11824 (2020)
20. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 1528–1540 (2016)