CONTEXTUAL ADVERSARIAL ATTACK AGAINST AERIAL DETECTION IN THE PHYSICAL WORLD

Jiawei Lian[1], Xiaofei Wang[1], Yuru Su[1], Mingyang Ma[2], Shaohui Mei[1]*

[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China
[2]School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, China

ABSTRACT

Deep Neural Networks (DNNs) have been extensively utilized in aerial detection. However, DNNs' sensitivity and vulnerability to maliciously elaborated adversarial examples have progressively garnered attention. Recently, physical attacks have gradually become a hot issue due to they are more practical in the real world, which poses great threats to some security-critical applications. In this paper, we take the first attempt to perform physical attacks in contextual form against aerial detection in the physical world. We propose an innovative contextual attack method against aerial detection in real scenarios, which achieves powerful attack performance and transfers well between various aerial object detectors without smearing or blocking the interested objects to hide. Based on the findings that the targets' contextual information plays an important role in aerial detection by observing the detectors' attention maps, we propose to make full use of the contextual area of the interested targets to elaborate contextual perturbations for the uncovered attacks in real scenarios. Extensive proportionally scaled experiments are conducted to evaluate the effectiveness of the proposed contextual attack method, which demonstrates the proposed method's superiority in both attack efficacy and physical practicality.

Index Terms— Adversarial examples, contextual perturbations, physical attacks, aerial detection

1. INTRODUCTION

In recent years, deep learning technology based on Deep Neural Networks (DNNs) has made great breakthroughs in computer vision, and natural language processing. Thus, DNNs have been widely applied in business and industry, such as mobile payment, autonomous driving, medical diagnosis, intelligent security, robotics, and other fields.

However, the widespread application of DNNs also buries potential safety hazards. Szegedy *et al.* [1] first designed an adversarial perturbation imperceptible to humans and added it to clean images to generate adversarial examples, which

can misguide DNNs make completely different wrong predictions. Such malicious behavior and maliciously designed examples are named adversarial attacks and adversarial examples, respectively, and the attacked model is also called the victim model. Since then, various deep learning tasks have fallen under adversarial attacks, such as image classification, object detection, spam detection, malware identification, natural language processing, deep reinforcement learning, *etc*. All DNNs-based models show great sensitivity and vulnerability in the face of adversarial examples.

Computer vision tasks, according to different attack domains, can be divided into digital attacks and physical attacks. Digital attacks refer to attack by tampering with the image pixels in the digital domain after imaging, while physical attacks refer to attack by tampering with the interested targets before imaging. Digital attack methods can easily fool various deep learning models in the digital domain. Since the generated digital perturbations typically cover the entire image and are invisible to humans, making them uncapturable by imaging devices. This problem drives more scholars to delve into the adversarial attacks applicable to real scenarios. Consequently, many physical attacks in patch form have been proposed to deceive intelligent systems such as autonomous driving [2], face recognition [3], and aerial detection [4] in real-world scenarios.

In this work, we devote ourselves to conducting contextual attacks (CA) against aerial detection in physical world scenarios. The main contributions are summarized as follows:

- We propose a novel contextual attack against aerial detection in physical scenarios, which achieves strong attack efficacy in both white-box and black-box conditions without smudging or blocking the targets to hide.

- We find that the targets' context information plays a key role in detection by observing their attention maps. Thus we make full use of the contextual feature of the interested targets to elaborate contextual perturbations.

- We evaluate the proposed contextual attack method with two SOTA methods by performing proportionally scaled experiments, demonstrating our method's superiority in both attack efficacy and physical robustness.

Fig. 1: The illustration of the proposed contextual attack against aerial detection in physical scenarios.

2. METHODOLOGY

2.1. Problem Formulation

In this work, we aim to design contextual perturbations in patch form for unblocked attacks in the physical world. Given a clean example x, by attaching the contextual perturbations on clean image x to generate adversarial example x^*. Technically, the adversarial example is formulated as follows:

$$x^* = (1 - M_{P^*}) \odot x + M_{P^*} \odot P^*, \qquad (1)$$

where $\odot$ and P^* represent Hadamard product and contextual perturbations, respectively. Perturbations' mask M_{P^*} is applied to properly attach contextual perturbations on the interested targets of a benign example.

2.2. Overall Framework

The overall framework of the proposed contextual attack is shown in Fig. 1. In the physical world, normal scenarios are captured by various remote sensing devices, such as satellites, aircraft, and drones. Then, various DNNs-based aerial detectors are adopted to process massive aerial imagery. To better understand the DNNs' predictions, we use Grad-CAM [5] to visualize the aerial detectors' attention maps. It is observed that the detectors also focus on the targets' contextual area beside the targets themselves. Based on the findings that the targets' contextual information plays an important role in aerial detection by observing the aerial detectors' attention maps. Thus we propose to fully use the contextual area of the interested targets to elaborate contextual perturbations for the uncovered attacks in real scenarios. Finally, the elaborately designed contextual perturbations are applied in the physical world to hide targeted objects from being recognized.

Fig. 2: Contextual perturbation design.

2.3. Contextual Attacks

Our proposed contextual attacks framework is mainly inspired by the following observations:

- Contextual features matters in aerial detection.

- Bigger contextual perturbations with stronger attack performance.

- Closer distance between perturbations and targets with stronger attack performance.

Therefore, to achieve powerful uncovered attacks in physical scenarios, we propose to manipulate the contextual area of the interested targets to elaborate contextual perturbations.

Specifically, we first extract the masks of the foreground M_{fg} and background M_{bg} of the targeted object T. Secondly, contextual perturbation's foreground F and background B area are extracted from the interested target T and updated perturbation P respectively. Thirdly, F and B are combined to formulate the contextual perturbation. Finally, repeat the above steps until the end of training. Mathematically, contextual perturbation is defined as:

$$\begin{aligned} P^c &= T \odot M_{fg} + P \odot M_{bg} \\ &= F + B \end{aligned} \qquad (2)$$

Table 1: Quantitative experimental results of white-box attacks in the physical world.

ADs	PAs	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	Avg
AD1	Clean	0.92	0.91	0.90	0.89	0.91	0.89	0.89	0.89	0.91	0.91	0.92	0.90	0.92	0.91	0.89	0.91	0.90	0.90	0.904
	PA1	0.82	0.77	0.75	0.50	0.83	0.00	0.84	0.84	0.87	0.73	0.80	0.86	0.82	0.82	0.60	0.83	0.87	0.87	0.746
	PA2	0.00	0.81	0.62	0.62	0.88	0.00	0.82	0.85	0.85	0.82	0.87	0.84	0.77	0.76	0.61	0.83	0.86	0.86	0.704
	Ours	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.000**
AD2	Clean	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.996
	PA1	**0.00**	0.99	0.97	0.21	0.37	0.00	1.00	1.00	0.98	0.76	0.95	1.00	0.98	**0.36**	0.96	1.00	1.00	1.00	0.752
	PA2	0.00	0.90	0.00	0.00	0.00	0.00	0.99	0.99	0.99	0.00	0.66	1.00	0.55	0.72	0.27	1.00	0.00	0.99	0.503
	Ours	0.27	**0.33**	**0.00**	0.93	0.99	**0.00**	**0.00**	**0.00**	**0.00**	0.32	0.98	**0.00**	**0.00**	0.93	**0.00**	**0.00**	**0.00**	**0.00**	**0.264**
AD3	Clean	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	0.998
	PA1	0.97	1.00	1.00	0.95	1.00	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.99	1.00	1.00	0.982
	PA2	0.98	1.00	0.93	0.25	1.00	0.74	0.99	1.00	1.00	0.84	0.94	0.97	0.97	1.00	0.00	1.00	1.00	1.00	0.867
	Ours	**0.00**	0.40	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.022**
AD4	Clean	1.00	1.00	1.00	1.00	1.00	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.988
	PA1	0.22	0.89	1.00	0.47	0.91	0.00	1.00	1.00	1.00	0.97	1.00	1.00	0.98	1.00	0.49	0.99	0.99	0.97	0.827
	PA2	1.00	0.91	1.00	0.82	0.94	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.43	0.99	1.00	1.00	0.894
	Ours	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.000**

The best attack performance are highlighted in **bold**.

Table 2: Quantitative experimental results of black-box attacks in the physical world.

ADs	PAs	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	Avg
AD2	PA1	0.99	0.94	0.99	0.89	0.78	0.00	1.00	1.00	1.00	0.95	0.99	1.00	0.96	0.63	0.98	1.00	0.99	0.99	0.893
	PA2	0.00	0.99	0.00	**0.00**	0.82	0.00	0.99	1.00	0.98	**0.44**	0.92	1.00	0.94	**0.25**	0.33	0.99	0.54	0.99	0.609
	Ours	**0.00**	**0.24**	**0.00**	0.51	**0.27**	**0.00**	**0.00**	**0.00**	**0.00**	0.80	**0.00**	**0.00**	**0.00**	0.49	**0.00**	**0.00**	**0.00**	**0.00**	**0.128**
AD3	PA1	1.00	1.00	1.00	0.99	1.00	0.42	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.967
	PA2	0.98	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.993
	Ours	**0.00**	**0.33**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.20	**0.00**	**0.00**	**0.00**	**0.00**	**0.029**
AD4	PA1	1.00	0.99	1.00	0.94	0.73	0.00	1.00	1.00	1.00	0.99	1.00	1.00	0.85	0.99	0.99	1.00	0.98	1.00	0.914
	PA2	0.98	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.85	1.00	1.00	0.96	1.00	0.88	1.00	0.92	0.99	0.921
	Ours	**0.00**	**0.00**	**0.00**	0.40	0.21	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.034**

The best attack performance are highlighted in **bold**. The proxy model is YOLOv5 (AD1).

2.4. Loss Function

To fool detectors in physical scenarios, the loss function consists of two components, adversary loss and smoothness loss.

Adversary loss: We use objectiveness scores of all detected objects to optimize the contextual perturbations. Therefore, the adversarial loss is written as:

$$L_{adv} = \frac{1}{n} \sum_{i=1}^{n} P_i(obj), \tag{3}$$

where $P_i(obj)$ means the objectiveness score of ith detected interested target , and n represents the number of detected interested targets. The adversarial loss is adopted to gift the contextual perturbations with attack efficacy during training.

Smoothness loss: Existing works demonstrate that perturbation's smoothness is crucial in maintaining attack efficacy during imaging. Since imaging devices can barely capture the value gap between adjacent pixels, total variation (TV) [6] is adopted as the smoothness limitation of the generated adversarial perturbations. TV can be written as:

$$L_{tv} = \sum_{i,j} \sqrt{(p_{i+1,j} - p_{i,j})^2 + (p_{i,j+1} - p_{i,j})^2}, \tag{4}$$

where $p_{i,j}$ represents the pixel value of ith row, jth column of the optimized adversarial perturbation.

Consequently, the total loss function is as follows:

$$L = L_{adv} + \lambda \cdot L_{tv}, \tag{5}$$

where λ is used to balance the two parts of the total loss.

3. EXPERIMENTS

3.1. Experimental Settings

In experiments, public datasets DOTA [7] and RSOD[1] are used to train aerial detectors (ADs) and contextual perturbations, respectively. Moreover, we choose various object detection methods to verify the attack effectiveness of the proposed method, including YOLOv5[2] (**AD1**), Faster R-CNN [8] (**AD2**), Swin Transformer [9] (**AD3**), FreeAnchor [10] (**AD4**). Two SOTA physical attacks (PAs) are chosen for comparison, including the adversarial perturbations generated by Thys *et al.* [11] (**PA1**) and APPA [4] (**PA2**).

(a) YOLOv5	(b) Faster R-CNN
(c) Swin Transformer	(d) FreeAnchor

Fig. 3: Qualitative attack performance in the physical world.

3.2. Experimental Results

The planes are chosen as the interested targets in the proportionally scaled experiments. We use the detection confidence scores of 18 plane models (**P1-P18**) to compare the physical attack performance, *i.e.*, the lower the confidence scores, the better the attack performance.

The quantitative experimental results of white-box and black-box attacks (YOLOv5 is selected as a proxy model to train contextual perturbations) are shown in Table 1 and Table 2 respectively. It is observed that our method shows great superiority in both attack efficacy and transferability. Specifically, our elaborated contextual perturbations can significantly lower the average detection confidence of all aerial detectors, from 0.904, 0.996, 0.998, 0.988 to 0.000, 0.264, 0.022, 0.000, far better than the comparison methods. Moreover, our method can drop the average detection confidence from 0.893, 0.967, 0.914 to 0.128, 0.029, 0.034, even in black-box settings. The qualitative experimental results are shown in Fig 3. We can observe that the generated contextual perturbations of our proposed contextual attack method can easily blind various aerial detectors, even after digital-physical domain transformation.

4. CONCLUSION

In this article, we aim to hide interested targets from being detected by various aerial detectors without smearing targeted objects. To achieve that, we propose a novel contextual attack against aerial detection in physical world scenarios, which fully uses the interested targets' contextual features to elaborate contextual perturbations and achieves the best attack performance in both white-box and black-box settings. Extensive experiments demonstrate the effectiveness and superiority of our proposed contextual attack method.

5. REFERENCES

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014. 1

[2] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu, "Dual attention suppression attack: Generate adversarial camouflage in physical world," in *CVPR*, 2021, pp. 8565–8574. 1

[3] Xingxing Wei, Ying Guo, and Jie Yu, "Adversarial sticker: A stealthy attack method in the physical world," *IEEE TPAMI*, 2022. 1

[4] Jiawei Lian, Shaohui Mei, Shun Zhang, and Mingyang Ma, "Benchmarking adversarial patch against aerial detection," *IEEE TGRS*, vol. 60, pp. 1–16, 2022. 1, 3

[5] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626. 2

[6] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *CCS*, 2016, pp. 1528–1540. 3

[7] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *CVPR*, 2018, pp. 3974–3983. 3

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, vol. 28. 3

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10012–10022. 3

[10] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye, "Freeanchor: Learning to match anchors for visual object detection," in *NIPS*, 2019, vol. 32. 3

[11] Simen Thys, Wiebe Van Ranst, and Toon Goedemé, "Fooling automated surveillance cameras: adversarial patches to attack person detection," in *CVPR*, 2019, pp. 0–0. 3