Physical Adversarial Attacks in the Era of AIoT

Tao Ni
Department of Computer Science
City University of Hong Kong
taoni2@cityu.edu.hk

***Abstract*—Artificial Intelligence of Things (AIoT) systems are increasingly deployed in applications such as autonomous driving and critical surveillance. However, the AI models powering these systems—particularly deep neural networks—are vulnerable to physical adversarial attacks through modalities like patches, projections, and infrared signals. In this paper, we present two physical adversarial attacks targeting traffic sign recognition (TSR) and face recognition (FR) systems. By manipulating physical signals, adversaries can induce dodging, denial-of-service, or impersonation attacks, potentially leading to serious consequences such as traffic accidents or identity theft.**

***Index Terms*—Physical adversarial attacks, AIoT, Traffic sign recognition, Face recognition**

I. Introduction

Recent years have witnessed the explosive development of AIoT systems, as well as their broad applications in various industries. However, recent studies [1], [2] have revealed that these AIoT systems are vulnerable to adversarial examples induced by physical signals, which are crafted inputs designed to mislead the AI models into making incorrect predictions or recognitions. To demonstrate the physical adversarial attacks in AIoT systems and raise public awareness, we introduce two physical adversarial attacks, named FIPatch and UVHat, targeting traffic sign recognition (TSR) and face recognition (FR) systems, respectively. By manipulating physical signals (*e.g.*, fluorescent ink, ultraviolet light) to create patches or interference, adversaries can launch dodging, denial-of-service (DoS), or impersonation attacks, which can potentially lead to serious consequences such as traffic accidents or identity theft. We have evaluated the proposed attacks in various real-world scenarios, where extensive experiments demonstrate the high effectiveness and practicality.

II. Threat Model

A. Adversary's Capabilities

Given the black-box settings of AI models deployed in various AIoT systems, we assume that the adversary has no direct access to internal details of the target model, such as its architecture, parameters, weights, or training data. The only available feedback is the output confidence scores for a given input. In practice, the adversary can query the model multiple times to obtain these output probabilities and iteratively fine-tune physical signals to launch physical adversarial attacks. In addition, we consider a budget for the number of queries (*i.e.*, less than 100 times) to carry out each attack due to the constraint on computational resources that the adversary can obtain in real-world scenarios, following the research line in relevant studies [3], [4].

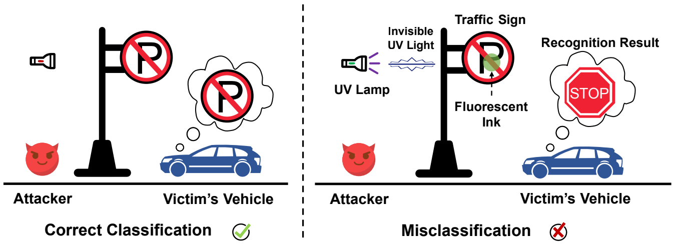

Fig. 1: An illustration of FIPatch attack.

B. Adversary's Goals

We consider three types of physical adversarial attacks: *(i) Dodging Attacks:* The AI models deployed in the AIoT system misidentified the input data in a given class as a different class, *i.e.*, an attacker on a blacklist being misclassified as someone off the list, thereby gaining unauthorized access to restricted areas such as banks or private residence; *(ii) Denial-of-Service (DoS) Attacks:* The adversary causes the AI models in the AIoT system to fail to detect any objects, which further disrupting the operations; and *(iii) Impersonation Attacks:* The adversary's profile is not in the database, but the AI models misidentified the adversary as other authorized identity in the database, *i.e.*, an adversary impersonating an authorized employee to gain the access permissions of a critical company to steal secret documents.

III. Case 1: Physical Adversarial Attacks on Traffic Sign Recognition (TSR) Systems

Traffic sign recognition (TSR) plays a pivotal role in autonomous driving by visually detecting and classifying traffic signs to ensure driving safety under various road situations. However, most TSR systems were built atop machine-learning models that are inherently suspicious and also shown to be subject to adversarial attacks [5], [6], which result in misidentification and severe traffic accidents.

We propose FIPatch [1], a stealthy physical adversarial patch attack using fluorescent ink, which exploits UV-triggered fluorescence to mislead traffic sign recognition (TSR) systems. As shown in Figure 1, the attacker applies carefully crafted fluorescent ink to a traffic sign, which becomes active under invisible UV light, leading to potential misclassifications. FIPatch consists of four modules: (1) a color-edge fusion method to localize traffic signs for precise ink application; (2) a simulation model capturing key fluorescent parameters

TABLE I: FIPatch evaluation results: The ASRs on various models in the physical world.

Ambient	Frames	Dodging Attack		DoS Attack		Impersonation Attack			
Light (Lux)		Yolov3	Faster R-CNN	Yolov3	Faster R-CNN	ResNet50	VGG13	MobileNet v2	GoogleNet
200	4374	98.31%	98.66%	91.59%	87.81%	100%	99.82%	98.93%	100%
500	3655	98.72%	93.41%	83.64%	80.09%	99.35%	98.01%	95.38%	97.52%
1000	4163	95.22%	88.31%	69.19%	64.91%	94.26%	92.58%	92.15%	93.66%
2000	3719	94.06%	86.10%	53.81%	48.43%	90.59%	86.37%	85.92%	84.03%
3000	3924	89.63%	83.39%	31.48%	25.92%	84.12%	79.55%	74.59%	76.15%

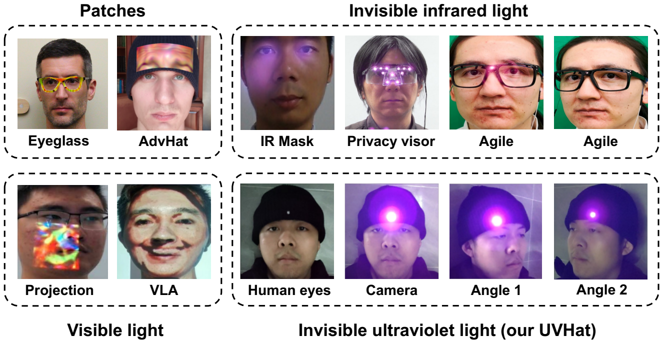

Fig. 2: Physical adversarial attack on FR systems.

(color, intensity, size); (3) goal-based, patch-aware loss functions targeting three attack goals; and (4) fluorescence-specific transformations to enhance real-world robustness.

We have evaluated FIPatch's effectiveness by testing the attack success rates (ASRs) on various conditions in the physical world. As shown in Table I, Yolov3 exhibits higher ASR compared to Faster R-CNN for both generative and hiding attacks, suggesting that Faster R-CNN is more robust against FIPatch. Dodging attacks are more effective than DoS attacks at all ambient light levels, while hiding attacks maintain ASRs above 80% only when the light is below 500 lux. At 3000 lux, the ASR for hiding attacks drops to below 32%. This drop is likely because detectors are more sensitive to perturbations on blank signs, such as contours, but more resilient when predicting existing traffic signs. All four classification models are highly vulnerable to attacks, with ASRs above 93% when light is below 1000 lux. Overall, the ASRs decreases with increasing ambient light because the fluorescent effect diminishes, reducing its impact on the models.

IV. Case 2: Physical Adversarial Attacks on AI-driven Critical Surveillance Systems

AI models, such as deep learning neural networks, which are employed in face recognition (FR) systems, have been shown to be vulnerable to physical adversarial attacks through various modalities, including patches, projections, and infrared radiation. However, existing adversarial examples targeting FR systems often suffer from issues such as conspicuousness, limited effectiveness, and insufficient robustness.

TABLE II: The ASR of UVHat at varying distances for DoS attacks.

Model	25cm	30cm	35cm	40cm	45cm	50cm
ArcFace	78%	73%	72%	70%	67%	54%
FaceNet	91%	92%	89%	80%	74%	65%
CosFace	83%	80%	78%	79%	69%	59%
MobileFace	79%	80%	79%	73%	66%	58%

As shown in Figure 2, we introduce a novel physical adversarial attack, UVHat, which utilises ultraviolet (UV) light emitted from a hat to disrupt FR models. In contrast to prior approaches, the primary challenges in developing UVHat lie in accurately simulating UV light sources on a curved surface and determining the optimal attack parameters in a black-box setting. Our method can be conceptualized as a three-step process. First, we devise an interpolation-based UV simulation technique that leverages a video interpolation model to generate UV images under varying distances, powers, and wavelengths within the digital domain. Second, we introduce a hemispherical UV modeling strategy to update the relevant parameters based on the positions across the curved surface. Finally, we employ a reinforcement learning optimization approach, wherein the agent iteratively explores the parameter space to identify the most effective attack parameters.

To evaluate the impact of the distance between the person and the camera on UVHat's performance, we conduct DoS attacks while keeping other attack parameters constant. The results shown in Table II indicate that as the distance increases, the ASR gradually decreases. Specifically, when the distance exceeds 45cm, there is a sharp decline in ASR. This is due to the nature of UV light, which disperses outward and is easily absorbed by the atmosphere. As the distance increases, the adversarial perturbation captured by the camera becomes weaker.

V. Conclusion

We introduce two physical adversarial attacks that target on common AIoT systems, including TSR systems in autonomous driving and FR systems in critical surveillance systems. The empirical results show that AI models deployed in these infrastructures are vulnerable to adversarial perturbations induced by external physical obfuscation or interference. We will focus on designing and developing effective defense approaches to build secure, reliable, and efficient AIoT systems.

References

[1] S. Yuan, H. Li, X. Han, G. Xu, W. Jiang, T. Ni, Q. Zhao, and Y. Fang, "ITPatch: An invisible and triggered physical adversarial patch against traffic sign recognition," *arXiv preprint arXiv:2409.12394*, 2024.

[2] S. Yuan, H. Li, R. Zhang, H. Cao, W. Jiang, T. Ni, W. Fan, Q. Zhao, and G. Xu, "Omni-angle assault: An invisible and powerful physical adversarial attack on face recognition," in *Proc. of ICML*, 2025.

[3] F. Croce, M. Andriushchenko, N. D. Singh, N. Flammarion, and M. Hein, "Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks," in *Proc. of AAAI*, 2022.

[4] G. Tao, S. An, S. Cheng, G. Shen, and X. Zhang, "Hard-label black-box universal adversarial patch attack," in *Proc. of USENIX Security Symposium*, 2023.

[5] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *NeurIPS*, 2019.

[6] C. Zhang, P. Benz, T. Imtiaz, and I. S. Kweon, "Understanding adversarial examples from the mutual influence of images and perturbations," in *Proc. of CVPR*, 2020.