Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

A comprehensive survey of physical adversarial vulnerabilities in autonomous driving systems*#

Shuai ZHAO[1,2,3], Boyuan ZHANG[1,2], Yucheng SHI[1,2], Yang ZHAI[1,2,3], Yahong HAN[†‡1,2], Qinghua HU[1,2]

[1]*College of Intelligence and Computing, Tianjin University, Tianjin 300072, China*
[2]*Tianjin Key Lab of Machine Learning, Tianjin 300072, China*
[3]*CATARC Intelligent and Connected Technology Co., Ltd., Tianjin 300000, China*
[†]E-mail: yahong@tju.edu.cn

Received Dec. 25, 2023; Revision accepted Apr. 7, 2024; Crosschecked Mar. 17, 2025

Abstract: Autonomous driving systems (ADSs) have attracted wide attention in the machine learning communities. With the help of deep neural networks (DNNs), ADSs have shown both satisfactory performance under significant uncertainties in the environment and the ability to compensate for system failures without external intervention. However, the vulnerability of ADSs has raised concerns since DNNs have been proven vulnerable to adversarial attacks. In this paper, we present a comprehensive survey of current physical adversarial vulnerabilities in ADSs. We first divide the physical adversarial attack methods and defense methods by their restrictions of deployment into three scenarios: the real-world, simulator-based, and digital-world scenarios. Then, we consider the adversarial vulnerabilities that focus on various sensors in ADSs and separate them as camera-based, light detection and ranging (LiDAR) based, and multifusion-based attacks. Subsequently, we divide the attack tasks by traffic elements. For the physical defenses, we establish the taxonomy with reference to input image preprocessing, adversarial example detection, and model enhancement for the DNN models to achieve full coverage of the adversarial defenses. Based on the above survey, we finally discuss the challenges in this research field and provide further outlook on future directions.

Key words: Physical adversarial attacks; Physical adversarial defenses; Artificial intelligence safety; Deep learning; Autonomous driving system; Data-fusion; Adversarial vulnerability
https://doi.org/10.1631/FITEE.2300867 **CLC number:** TP391

1 Introduction

Autonomous driving (AD) systems (ADSs) constitute multiple perception-level tasks that have achieved high precision because of deep learning architectures (Kiran et al., 2022). With the help of artificial intelligence (AI) based self-driving architectures, ADSs can perceive various traffic assignments. To handle real-world scenes through visual complexity, ADSs deploy multiple sensors such as light detection and ranging (LiDAR), radar, and global positioning system (GPS) to assist camera sensors. Moreover, the growing interest in the development of ADSs has introduced new security challenges and vulnerabilities. Besides the usual cyber-attacks (Dibaei et al., 2020), such as denial-of-service (DoS) attack, black-hole attack, and malware attack, the vulnerability of deep neural networks (DNNs) in ADSs needs to be investigated. DNN predictions can be manipulated by adversarial examples

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (Nos. 62376186 and 61932009)
# Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2300867) contains supplementary materials, which are available to authorized users
ORCID: Shuai ZHAO, https://orcid.org/0000-0002-8745-9433; Yahong HAN, https://orcid.org/0000-0003-2768-1398

(Szegedy et al., 2014). In the early stage, most adversarial examples are designed in the digital world, which means that the adversarial perturbations are generated in a pixel-to-pixel manner. These digital adversarial samples cannot effectively attack ADSs. Research indicates that digital adversarial attacks perform poorly against ADSs due to two potential reasons: real-world properties and multisensor fusions (MSFs) (Boloor et al., 2019).

Things have changed since Eykholt et al. (2018b) proposed the physical adversarial attack for object detectors. They evaluated the difference between digital and physical adversarial attacks from four perspectives: environmental conditions, spatial restrictions, physical limits on imperceptibility, and fabrication error. The environmental conditions represent the camera in autonomous vehicles taking photos of adversarial examples from diverse angles, distances, and weather conditions. Attackers can manipulate only the "restrictive region" rather than background imagery. The physical limits on imperceptibility impose that physical adversarial perturbations should be stealthy instead of imperceptible so that the camera can perceive perturbations. The fabrication error means that all perturbation values must be valid colors that can be reproduced in the real world. Furthermore, even if a fabrication device such as a printer can produce specific colors, some reproduction errors will occur in the progress of digital-to-physical transformation (Fig. 1).

Sensor fusion involves gathering inputs from multiple sensors to interpret environmental conditions with increased detection certainty. The integration of diverse sensor types in ADSs allows them to harness the collective advantages of the sensors, effectively compensating for their limitations. Therefore, the majority of today's automotive manufacturers commonly use the following sensors in ADSs: visible camera, LiDAR, infrared camera, depth camera, GPS, radar, and ultrasound. Instead of digital adversarial attacks attacking mainly the vanilla images' visual information, physical adversarial attacks consider fooling multiple sensors in ADSs simultaneously.

To address the above issues, we propose a novel taxonomy to map the ADSs under adversarial attacks and defenses. We analyze the adversarial attacks of ADSs from three perspectives: attack scenarios, attack sensors, and attack tasks. After that, we investigate the robustness of ADSs and discuss the corresponding defense strategies in the following three steps: input image preprocessing, adversarial example detection, and model enhancement. The whole structure is shown in Fig. 2.

Owing to the size of the adversarial threat, we

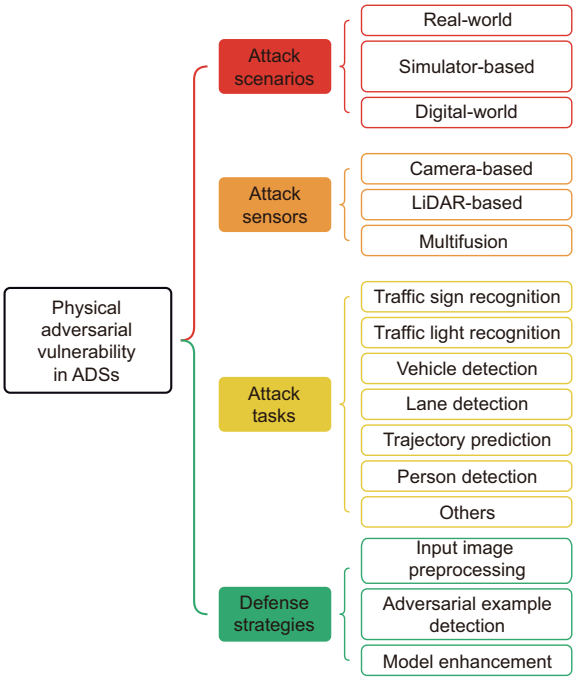

Fig. 2 Framework of this survey. We analyze the adversarial vulnerability from four perspectives: attack scenarios, attack sensors, attack tasks, and defense strategies

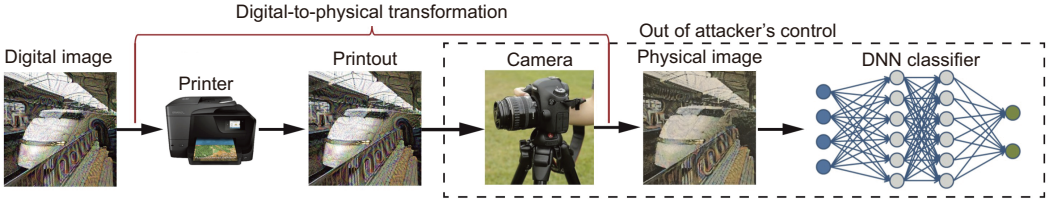

Fig. 1 Adversarial examples are transformed across the digital and physical worlds before they enter the DNN image classifier. In practice, attackers have no (limited) control over the internal system (Jan et al., 2019)

classify attack scenarios into three scenarios: real-world, simulator-based, and digital-world ones. The real-world scenario represents attackers making up the physical adversarial examples in the real world, and we evaluate their effectiveness using a few natural ADSs and object detectors. These physical adversarial examples present significant threat, harm, and risk to ADSs. The simulator-based scenario represents the attackers generating physical adversarial examples in simulators by mimicking various elements from the real world; the adversarial examples in simulator-based scenarios present threats because they usually cause simulation-to-real scene transfer issues. The digital-world scenario represents the attackers designing adversarial examples against ADSs, but they evaluate their methods only in the digital world. Their method can inspire follow-up research on the adversarial vulnerability of ADSs.

Subsequently, we categorize the attack sensors into camera-based attacks, LiDAR-based attacks, and multifusion attacks. Within the realm of camera-based attacks, adversarial examples are designed exclusively to deceive the visible camera sensor. LiDAR-based attacks are tailored to produce detrimental adversarial examples against LiDAR perception systems. Multifusion attacks encompass scenarios wherein multiple sensors are concurrently targeted. In this context, we explore infrared, depth-based, and radar-based attacks, as these sensors typically operate in conjunction rather than in isolation.

In each sensor, we consider the adversarial vulnerability from different attack tasks, i.e., traffic sign recognition, traffic light recognition, vehicle detection, lane detection, trajectory prediction, person detection, and others. In each task, we compare the adversarial threats between various adversarial examples.

After that, we analyze the adversarial robustness of ADSs and discuss recent defense strategies that target enhancing adversarial robustness considering three subcategories: input image preprocessing, adversarial example detection, and model enhancement. Each perspective and its corresponding examples are shown in Fig. 3.

Fig. 4 demonstrates the related publications in this research direction since 2017. In 2017–2019, research has mainly discussed the difference between digital and physical adversarial examples. These works made a few attempts to explore the adversarial vulnerability of ADSs. Since 2020, existing publications have began to design powerful adversarial attacks against ADSs and proposed corresponding defense strategies.

Although there are a few surveys about adversarial attacks and defenses in computer vision, they lack a comprehensive and latest focus on physical adversarial attacks and defenses in ADSs. For example, Serban et al. (2021) summarized adversarial attacks and defenses in object detection, and Machado et al. (2021) considered adversarial machine learning in the

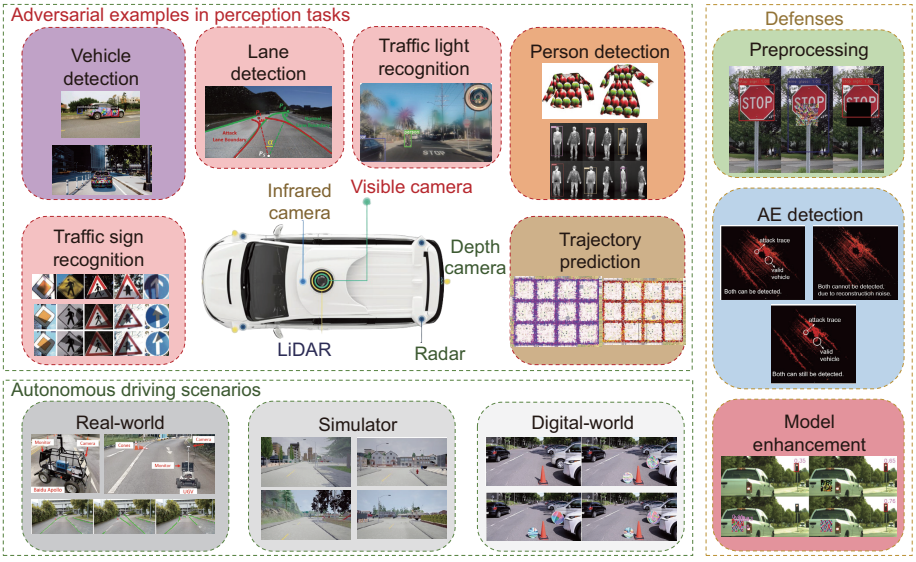

Fig. 3 Adversarial examples in autonomous driving systems and different perspectives to evaluate the vulnerability of ADSs. A few examples of recent research are presented to illustrate each subcategory

classification task. Qiu et al. (2019) introduced physical adversarial attacks as an application scenario of adversarial attacks. The above surveys in adversarial machine learning focus mainly on the digital domain while largely ignoring the physical world. Recently, Wei H et al. (2022) investigated the physical adversarial attack in computer vision. They evaluated the physical adversarial examples in terms of effectiveness, stealthiness, and robustness. Ansari and Singh (2021) investigated mainly the physical adversarial attacks of person detection tasks in real-time surveillance. Modas et al. (2020) briefly introduced the emerging field of sensing (ultrasonic, radar, GPS, LiDAR, and camera) for autonomous vehicles. Wei XX et al. (2022) analyzed visual adversarial attacks and defenses in the physical world. Ding et al. (2023) discussed the safety-critical driving scenario generation, which focuses mainly on related datasets and simulators. The current survey is different from the above surveys from three perspectives: First, previous surveys barely mentioned whether the adversarial examples have proven their real-world performance. Moreover, the current survey investigates the computer vision community's physical adversarial attacks and defenses, while the physical adversarial vulnerabilities in ADSs consider multiple sensors. Finally, we consider person detection as an essential perception task to analyze the adversarial vulnerability of ADSs, which was ignored by previous surveys.

The background knowledge is discussed in the supplementary materials.

2 Real-world scenario

Physical adversarial attacks are one of the essential ways to evaluate the adversarial vulnerability of ADSs. Current attackers have built different adversarial examples to spoof different perception tasks in ADSs. Table 1 summarizes the related physical adversarial attack methods. The corresponding proportion of attacks per perception task is shown in Fig. 5. The table and figure show that most physical adversarial attacks focus on traffic sign recognition and vehicle detection tasks, which account for 50% of all attacks. Person detection attacks take up a significant proportion of all adversarial threats in ADSs. The above three tasks take up about 64% of all physical attacks in ADSs. Furthermore, a few research works investigated the adversarial example against

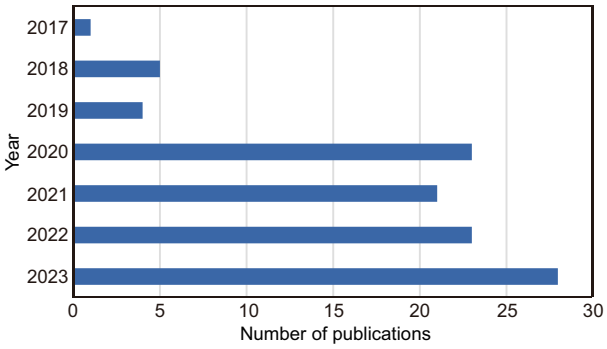

Fig. 4 Research related to physical adversarial machine learning in ADSs has grown since 2020

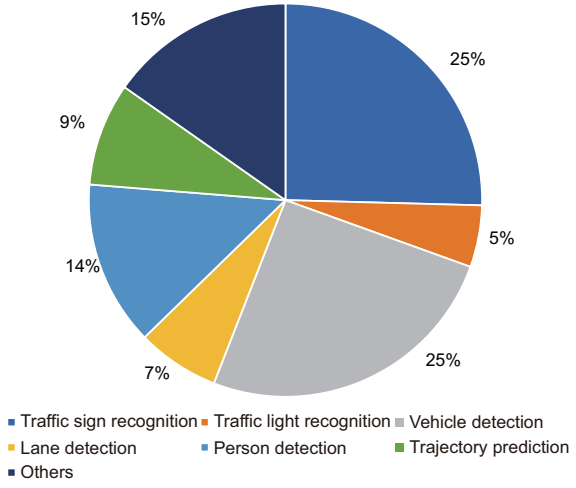

Fig. 5 Proportion of physical adversarial attacks per perception task

traffic light recognition, lane detection, and trajectory prediction tasks. The others represent several tasks that are barely talked about, such as monocular depth estimation (MDE) and license plate recognition (LPR). Besides attack scenarios, attack sensors, and attack tasks, we summarize the attack settings of all physical attacks. Most adversarial attacks are designed in the white-box setting and their performance is evaluated under the transfer-based black-box setting. The current physical adversarial examples do not link a particular design to enhance their transferability.

In this section, we first discuss the physical adversarial examples in real-world scenarios. Those methods have proven harmful to ADSs because they evaluate their effectiveness under a few natural ADSs and object detectors. We divide the physical adversarial examples into three categories: camera-based, LiDAR-based, and multifusion attacks. In each category, we will discuss them depending on their target perception tasks.

Table 1 Physical adversarial attack methods mentioned in this paper

Attack	Scenario			Sensor			Task								Setting		
	Real	Simulator	Digital	Camera	LiDAR	Others	TSR	TLR	PD	VD	TP	MDE	LD	Others	White	Black	
																Query-based	Transfer-based
Shadow attack (Zhong et al., 2022)	✓			✓			✓									✓	
TC-EGA (Hu ZH et al., 2022)	✓			✓					✓								✓
DTA (Suryanto et al., 2022)		✓		✓						✓					✓		
Zhang QZ et al. (2022)'s		✓		✓							✓				✓	✓	✓
RP2 (Eykholt et al., 2018b)	✓			✓			✓								✓		
FLAT (Li YM et al., 2021)			✓		✓						✓						✓
FCA (Wang DH et al., 2022)		✓		✓						✓							✓
DAS (Wang JK et al., 2021)		✓		✓						✓					✓		
CAMOU (Zhang Y et al., 2019)		✓		✓						✓						✓	
Zhu XP et al. (2022)'s	✓					✓			✓								✓
AdvDO (Cao et al., 2022)			✓	✓							✓				✓		
Cheng et al. (2022)'s	✓			✓		✓						✓					✓
RP2D (Eykholt et al., 2018a)	✓			✓			✓								✓		
AdvCam (Duan et al., 2020)	✓			✓			✓								✓		
TAA (Yang XH et al., 2021)	✓			✓			✓								✓		
Zolfi et al. (2021)'s	✓			✓				✓									✓
B&S (Patel et al., 2020)		✓		✓				✓							✓		
ER (Wu T et al., 2020b)		✓		✓						✓						✓	
Abdelfattah et al. (2021a)'s			✓	✓	✓					✓					✓		
MSF-ADV (Cao et al., 2021)	✓			✓	✓									✓	✓		
Abdelfattah et al. (2021b)'s			✓	✓	✓					✓					✓		
Boloor et al. (2020)'s		✓		✓									✓			✓	
Han et al. (2022)'s	✓			✓									✓		✓		
DRP (Sato et al., 2021)	✓			✓									✓		✓		
Yamanaka et al. (2020)'s	✓					✓				✓					✓		
Jia et al. (2022)'s	✓			✓			✓										✓
Lu et al. (2017)'s	✓			✓			✓								✓		
ShapeShifter (Chen ST et al., 2019)	✓			✓			✓								✓		
NaturalAE (Xue et al., 2021)	✓			✓			✓								✓		✓
M-SimBA (Kumar et al., 2020)	✓			✓			✓									✓	
IAP (Bai et al., 2022)	✓			✓			✓										✓
Brand et al. (2022)'s		✓		✓							✓				✓		
PhysGAN (Kong et al., 2020)	✓			✓							✓						✓
Spot evasion attack (Qian et al., 2020)	✓			✓										✓			✓
Frustum attack* (Hallyburton et al., 2022)		✓		✓	✓					✓					✓		
fakeWeather (Marchisio et al., 2022)			✓	✓										✓			✓
Yoon et al. (2023)'s	✓			✓										✓	✓		
Adv-Plate (Kwon and Baek, 2021)	✓			✓										✓	✓		
Adv-LiDAR (Cao et al., 2019)		✓			✓					✓					✓		
Choi and Tian (2022)'s			✓	✓										✓	✓		
Chen J et al. (2022)'s			✓	✓										✓	✓		
a-RNA (Guesmi and Alouani, 2022)	✓					✓								✓	✓		
Rogue signs (Sitawarin et al., 2018)	✓			✓			✓								✓		
DT-UAPs (Benz et al., 2020)	✓			✓			✓								✓		
SLAP (Lovisotto et al., 2021)	✓			✓				✓							✓		
UPC (Huang LF et al., 2020)	✓			✓					✓						✓		
Wu ZX et al. (2020)'s	✓			✓					✓								✓
Hu YCT et al. (2021)'s	✓			✓					✓								✓
Zhu XP et al. (2021)'s	✓			✓					✓								✓
AdvLB (Duan et al., 2021)	✓			✓			✓									✓	
GradOpt (Yang JH et al., 2020)		✓		✓									✓		✓		
GhostImage (Man et al., 2020)	✓			✓			✓										✓
PRA (Cao et al., 2023)	✓				✓									✓			✓
Tu et al. (2020)'s		✓			✓					✓					✓	✓	
Yang KC et al. (2021)'s		✓			✓					✓					✓	✓	
Frustum attack** (Hallyburton et al., 2022)		✓		✓	✓					✓							✓
Tu et al. (2022)'s			✓	✓	✓					✓							✓
Thys et al. (2019)'s	✓			✓					✓						✓		
Adversarial T-shirt (Xu KD et al., 2020)	✓			✓					✓						✓		
Meta-GAN (Feng et al., 2024)			✓	✓			✓								✓		
Wei XX et al. (2023)'s	✓			✓		✓			✓						✓		✓
Rossolini et al. (2023)'s			✓	✓										✓			✓
3D consistent patch attack (Zhu ZJ et al., 2023)			✓	✓	✓					✓					✓		

Frustum attack has two different settings. TSR: traffic sign recognition; TLR: traffic light recognition; PD: person detection; VD: vehicle detection; TP: trajectory prediction; LD: lane detection

2.1 Camera-based attacks

Visible cameras play an important role in the perception of ADSs. Hence, it is necessary to investigate the adversarial vulnerability of camera sensors. In this subsection, we discuss mainly the vulnerability of visible cameras in safety-critical tasks related to ADSs, including traffic sign recognition, person detection, traffic light recognition, and lane detection.

1. Traffic sign recognition

Traffic sign recognition aims to enable autonomous vehicles to detect and interpret various traffic signs, such as speed limits, stop signs, yield signs, and warning signs. The system must accurately recognize and interpret these signs to make informed decisions while navigating the road. In the early stage, attacks on traffic signs are usually carried out against the classification models under the white-box setting. Lu et al. (2017) proposed an attack on traffic sign detectors, but their method is not robust enough and needs to add large perturbations to be successful. Autonomous vehicles take pictures of traffic sign recognition at different distances, angles, and weather conditions. Compared to digital attacks, enhancing the robustness of physical adversarial examples is necessary.

Eykholt et al. (2018b) observed that the main challenge in generating robust physical environments is the environmental variability, i.e., the distance and angle of the viewing camera. They proposed robust physical perturbations (RP2) to generate robust adversarial perturbations under traffic signs. The objective function of RP2 is as follows:

$$\begin{aligned}&\underset{\delta}{\operatorname{argmin}}\,||M_x\delta||_p+\mathrm{NPS}(M_x\delta)\\&+\mathbb{E}_{x_i\sim X^V}J(f_\theta(x_i+T_i(M_x\delta)),y^*).\end{aligned}\tag{1}$$

This can be divided into three parts. The first term controls the ℓ_p norm of the δ perturbation masked by M_x. The second one measures the non-printability score (NPS), and the last term represents the expected value of the loss function J. RP2 first uses NPS to measure the domain gap between the digital and real worlds. The authors also proposed a novel two-stage evaluation methodology to evaluate the robustness of adversarial examples. In the first laboratory stage test, the viewing camera is kept at various distances and angle configurations. The second field stage test involves the victim driving a car toward an intersection in uncontrolled conditions to simulate an autonomous vehicle. To reduce the attention of human eyes, they introduced only a set of black-and-white stickers as perturbations. Kumar et al. (2020) proposed M-SimBA to balance the adversarial transferability and performance. This approach helps overcome the late-convergence issue and produces adversarial samples with the least confidence in the ground-truth class.

Feng et al. (2024) reconceptualized physical attacks through the lens of few-shot learning, wherein the training task was reformulated to consist of a support set, a query set, and a designated target DNN model. To improve the generative attack model's generalization ability, they introduced the meta-generative adversarial network (GAN) attack, an approach that is both class-agnostic and model-agnostic, leveraging the principles of meta-learning. This method uses CycleGAN to reduce the need for manual labor. When benchmarked against contemporary leading physical attack techniques, the Meta-GAN attack demonstrates enhanced robustness and superior generalization capabilities.

The above methods have demonstrated the vulnerability of traffic sign recognition in ADSs. They usually generate image-dependent perturbations to fool ADSs for one specific instance, while universal adversarial perturbations (UAPs) have proven their capability to fool the perception module by adding one perturbation to any image in a dataset (Moosavi-Dezfooli et al., 2017; Chaubey et al., 2020). Benz et al. (2020) designed the double-targeted universal adversarial perturbations (DT-UAPs) to generate image-agnostic perturbations and make them suitable for real-time applications such as autonomous driving and robotics.

After investigating the traffic sign adversarial examples on image classification models, Eykholt et al. (2018a) extended the physical adversarial setting in RP2 to object detection models and testified it with traffic signs. According to the painting of adversarial patches on a stop sign, robust physical-world perturbation attacks on deep learning visual classification (RP2D) could make the sign disappear under object detectors. Two examples are illustrated in Fig. 6. Chen ST et al. (2019) designed ShapeShifter attack to enhance the expectation-over-transformation (EOT) technique and iteratively generated the adversarial perturbations. NaturalAE

(Xue et al., 2021) uses an adaptive mask to constrain the area and intensities of the perturbations, and a real-world perturbation score (RPS) is used to make the perturbations similar to the real noise in the physical world. Due to incorporation of various image transformation functions, the adversarial transferability and robustness are enhanced. Duan et al. (2020) observed that stealthiness and flexibility are more important in physical adversarial attacks. They designed style loss, content loss, and smoothness loss, and proposed adversarial camouflage (AdvCam) to transfer large adversarial perturbations into customized styles. Yang XH et al. (2021) proposed the targeted attention attack (TAA) based on soft attention maps. Experimental results validated the attack performance of TAA while reducing the perturbation loss. Therefore, these adversarial examples are easy to deploy in the real world. Jia et al. (2022) designed the S-BBOX and M-BBOX filters after modifying the four loss functions for the four attack vectors: hiding attack (HA), appearance attack (AA), nontarget attack (NTA), and target attack (TA). Besides the above methods evaluating the adversarial robustness on different distances and angles, the generated physical adversarial examples enhance the robustness under various environment illuminations.

Fig. 6 Two examples of the extended RP2 attack with poster and sticker attacks (Eykholt et al., 2018a)

The patch-based attack mentioned previously, while effective, may not go unnoticed in real-world scenarios. Such adversarial perturbations, typically printed on stickers and applied to the target object, could draw the attention of observant drivers and thus be readily identified. Moreover, these attacks encounter practical challenges, such as the attacker's physical access to the target and the accurate reproduction of colors that are faithful to the original design in varying real-world conditions. Zhong et al. (2022) categorized these types of physical-world adversarial attacks as "sticker-pasting" methods due to their application technique. In contrast, they developed an "optic-based" approach, referred to as the shadow attack, which offers a more subtle and possibly less-detectable method of executing physical-world adversarial actions. The phenomenon is shown in Fig. 7. Based on the phenomenon, they used shadow to propose a novel query-based attack, namely, shadow attack. Hence, the adversarial examples generated by the shadow attack are more naturalistic and stealthy.

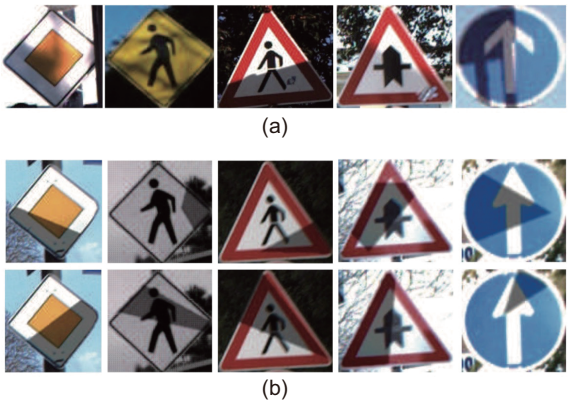

Fig. 7 The image in the panel shows the shadows in the real world (a). Following panel (a), Zhong et al. (2022) used the shadow attack to generate stealthy adversarial examples shown in panel (b)

Instead of shadow attacks, Lovisotto et al. (2021) proposed other types of "optic-based" attacks, such as short-lived adversarial perturbation (SLAP) which allows adversaries to use light projectors to project adversarial perturbations onto real-world objects. They provided attackers with greater control over the attack by allowing them to turn the projections on or off as needed. The SLAP-generated adversarial example is effective against existing defense mechanisms such as SentiNet (Chou et al., 2020), a state-of-the-art defense method against adversarial patches.

Duan et al. (2021) proposed the adversarial laser beam (AdvLB) attack to leverage a laser beam as an adversarial perturbation on traffic signs, which indicates two main causes of prediction errors of DNNs caused by AdvLB attack. First, the laser beam's color feature changes the raw image and forms a new cue for DNNs. Second, the laser beam introduces dominant features of specific classes, especially

lighting-related classes. The presence of the laser beam biases the DNN toward the feature introduced by the beam, resulting in misclassification.

The aforementioned traffic sign adversarial examples are designed by pasting patches and textures on existing traffic signs. Sitawarin et al. (2018) generated adversarial examples by modifying innocent signs and advertisements in the environment to be classified as the attacker's desired traffic sign with high confidence. Unlike previous attacks, rogue signs allow for the modification of any innocuous sign. The authors included various image transformations in the optimization problem to generate physically robust adversarial samples.

2. Person detection

Person detection is an essential task for ADSs. Physical adversarial attacks against person detectors might lead to vehicles not being able to detect pedestrians on the road, which leads to a personal security threat. Thys et al. (2019) attacked the person detector in ADSs with a white-box setting, where they used a printed version of their adversarial patch and held it in hand to evade pedestrian detection. Xu KD et al. (2020) designed the adversarial T-shirt for evading person detectors. They leveraged a thin plate spline (TPS) based Transformer to model the nonrigid deformation of the T-shirts.

Huang LF et al. (2020) pioneered the concept of attacking nonrigid or nonplanar objects through the creation of the universal physical camouflage (UPC) attack. Their evaluation spanned various detection tasks, including those for persons and vehicles. To improve the camouflage of the perturbations, they crafted a technique that projects patterns onto the surface of ℓ_∞-norm balls with a radius ϵ centered around a natural image I. The formulation is shown below:

$$\delta^t = \mathrm{Proj}_\infty(\delta^{t-1} + \Delta\delta, I, \epsilon). \tag{2}$$

To encourage the following research, they also proposed AttackScenes, a benchmark dataset that simulates the real three-dimensional (3D) world in a controllable and reproducible environment.

Wu ZX et al. (2020) considered the deformation of the nonrigid body and designed adversarial examples using printed posters and wearable clothing, taking into account various real-world factors, such as cameras, lighting conditions, distances, and angles. The result of this person evasion attack is demonstrated in Fig. 8. To produce natural adversarial patterns and make a person vanish under the object detector, Hu YCT et al. (2021) used a pre-trained BigGAN (Brock et al., 2018) to control specific category generation. By sampling images from these GANs that minimize the detection score of a target object, the method can generate natural-looking patches while maintaining acceptable attack performance.

Hu ZH et al. (2022) stated that the above patch-based adversarial examples (Adv-Patch) could have catastrophic drops in the attack success rate when the viewing angle (i.e., the camera's angle toward the object) changes. To enhance the multiview robustness of adversarial examples, they designed a method named adversarial texture (Adv-Texture). The benefits they claimed for Adv-Texture are shown in Fig. 9. More specifically, they proposed a toroidal-cropping-based expandable generative attack (TC-EGA), which is based on pretrained GANs to successfully generate more natural-looking adversarial patches. The experiments illustrated that TC-EGA has better attack performance and transferability at multiple viewing angles than Adv-Patches.

Fig. 8 In this demonstration, the person detector was evaded using a pattern trained on the Microsoft Common Objects in COntext (COCO) dataset with a carefully constructed objective (Wu ZX et al., 2020)

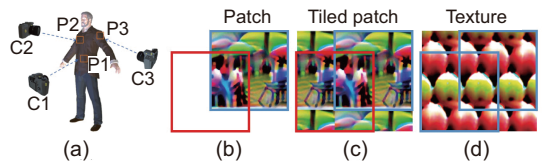

Fig. 9 Illustration of the attacks at different viewing angles (Hu ZH et al., 2022): (a) the camera captures different parts (P1, P2, and P3) of the clothes when set to different viewing angles (C1, C2, and C3); (b–d) the boxes are the possible areas that the camera may capture. The blue ones indicate the most effective areas for attack, while the red ones are less effective. References to color refer to the online version of this figure

3. Traffic light recognition

Traffic light recognition systems are vulnerable to physical adversarial attacks. Zolfi et al. (2021) designed contactless translucent physical patches and placed them on the camera's lens. By applying the patches on the front camera of the Tesla Model X, they carried out a targeted attack on Tesla autopilot.

4. Lane detection

Lane detection tasks would ensure that the vehicles do not deviate from the right road lane. Han et al. (2022) brought physical backdoor attacks in the lane detection system. They designed two attacks to attack the autonomous system. In the poison-annotation attack, the adversary intentionally mis-annotates samples with a specific trigger to craft poisoned samples. In the clean-annotation attack, they exploited the image-scaling vulnerability to create visually similar poisoned samples with incorrect annotations. The results showed its effectiveness on the Baidu Apollo D-Kit vehicle with a Leopard camera and Weston Unmanned Ground Vehicle with a RealSense D435i camera.

Sato et al. (2021) analyzed various transportation policy guidelines; they summarized that the average driver reaction time to road hazards is commonly considered to be around 2.5–3.0 s. The average driver reaction time includes the time when the driver's eyes see a hazard until the driver's brain recognizes it (average perception time) and the time from the driver's brain recognizing the hazard to physically taking actions (average reaction time). Depending on the above observation, they investigated the security of deep learning based automated lane centering (ALC) systems against physical-world adversarial attacks. They proposed a novel dirty road patch (DRP) approach to generate adversarial examples, which can successfully attack the ALC systems.

2.2 LiDAR-based attacks

In real-world scenarios, LiDARs have a significant impact on ADSs. LiDAR sensors play a critical role in perceiving and understanding the surrounding environment. They are responsible for detecting obstacles, measuring distances and velocities, and providing real-time environmental information to the ADS. LiDAR sensors cannot be affected by image-based adversarial examples; this leads to more robustness than camera sensors. However, recent research demonstrates that LiDAR sensors are vulnerable to potential adversarial examples generated by 3D mesh and laser, which will affect trajectory prediction tasks. The adversarial examples against trajectory prediction exhibit two distinctive characteristics. First, the adversarial trajectories should obey physical rules and are possible to happen in the real world. Hence, the trajectories can be reproduced by the attacker-controlled vehicle in the real world and cannot be easily classified as anomalies by autonomous vehicles (AVs). Second, the attackers need to define optimization objectives that are semantically attractive for their targeting trajectory prediction. Kong et al. (2020) designed PhysGAN to produce imperceptible adversarial posters on the roadside board. To enhance the adversarial robustness and transferability, PhysGAN considers the background imagery typically captured by a camera during driving. They evaluated the physical adversarial effectiveness against NVIDIA Dave-2 on a real car.

2.3 Multifusion attacks

While the above camera-based and LiDAR-based attacks have shown the potential to mislead the ADSs into erroneous perceptions, contemporary ADSs increasingly rely on MSF modules. These modules integrate data from various sensors, such as cameras, LiDAR, radar, and GPS, to construct a cohesive and accurate representation of the vehicle's surroundings. This integration is crucial for making reliable decisions, especially in complex and dynamic driving environments.

However, as the ADS technology advances, the

strategies for adversarial attacks are also continuously evolving. Recognizing the reliance of ADS on sensor fusion for accurate perception, recent physical adversarial attacks have evolved to simultaneously target multiple sensors to exploit the data fusion process. These attacks are designed to create discrepancies in the data collected from different sensors, leading to incorrect predictions by the system. Those multifusion attacks could also spoof different perception tasks, such as vehicle detection, MDE, and person detection. However, due to the cost of evaluation of multifusion attacks in the real world, we mainly discuss them in the simulator-based scenarios and digital-world scenarios.

1. Vehicle detection

Cao et al. (2021) presented a study on the security of MSF based perception in ADSs. They introduced a novel method called MSF-ADV, which generates physically realizable, adversarial 3D-printed objects that mislead an ADS and cause accidents. The attack is formulated as an optimization problem. It addresses the challenges of nondifferentiable target cameras and LiDAR sensing systems and nondifferentiable cell-level aggregated features used by LiDAR. MSF-ADV has proved its performance in industry-grade ADS using real-world driving scenarios and a production-grade simulator.

2. MDE networks

Yamanaka et al. (2020) first attacked MDE networks with adversarial patches in white-box settings. They chose digital attacks to generate a patch and optimized the depth values in the patch region to certain depth values, which leads MDE networks to produce incorrect depth estimates. Cheng et al. (2022) proposed a natural-style patch against MDE models. They balanced the stealth and effectiveness by controlling the patch size with a differentiable mask. The experiments demonstrated their impact on downstream 3D reconstruction tasks.

3. Person detection

Zhu XP et al. (2022) developed infrared invisibility clothes composed of a combination of thermal insulation material (TIM) and regular fabric. When observed through a thermal camera, the area covered by TIM appears dark, while the remaining fabric appears bright. They simulated the cloth-to-clothing process in the digital world to fool infrared detectors from different angles. Consequently, this distinct thermal contrast can be exploited by adversaries to create a black-and-white adversarial pattern that resembles a "quick response code."

Wei XX et al. (2023) took a novel approach to circumvent both visible and infrared pedestrian detection systems simultaneously by printing digitally simulated results and then cutting them out using insulating materials to create patches. They used a method that combines boundary-limited shape optimization with a score-aware iterative evaluation process. This technique was designed to generate effective patches in the digital realm that maintain a balance between evading detection by multimodal detectors. Zhu XP et al. (2021) proposed an adversarial light attack to craft adversarial examples with small bulbs on a board against the thermal infrared pedestrian detectors. They designed a transformation module to transform and test the bulb board on the YOLOv3 (You Only Look Once) model. The attack achieved satisfactory performance, which indicates the vulnerability of the infrared camera in ADSs. The whole structure is shown in Fig. 10.

3 Simulator-based scenario

In this section, we discuss the physical adversarial examples in the simulator-based scenario. These adversarial examples have evaluated their attack ability in traffic simulators. However, the adversarial examples in simulators usually cause simulation-to-real scene transfer issues. Hence, these physical adversarial attacks need further investigation to confirm their capability.

3.1 Camera-based attacks

1. Vehicle detection

Suryanto et al. (2022) claimed that existing neural tenderers have no capacity to fully represent various real-world transformations. Therefore, they proposed a differentiable transformation network (DTN) to learn the expected transformation of a rendered object. Then, they proposed the differentiable transformation attack (DTA) to camouflage the target vehicle against object detection models with a wide range of transformations.

2. Lane detection

Boloor et al. (2020) proposed a query-based method with a Bayesian optimization strategy against end-to-end self-driving models. The novel hijacking attack paints black lines on the road and

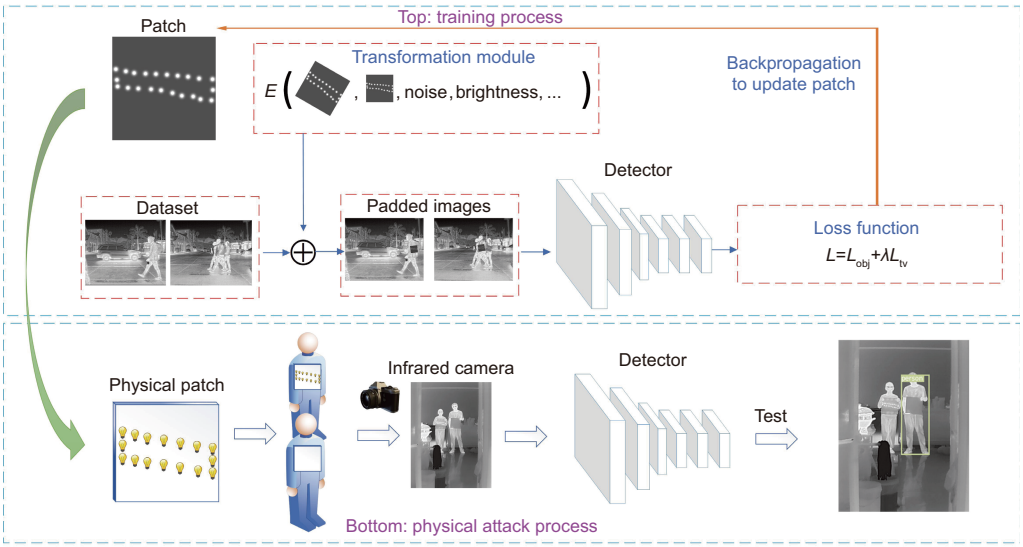

Fig. 10 Overview of Zhu XP et al. (2021)'s framework

causes the car to follow a target path, even when it is quite different from the correct route (e.g., causing the car to turn left instead of right). Yang JH et al. (2020) proposed GradOpt attack as a scalable approach to generated adversarial examples. In the CARLA simulator, they proved that GradOpt is available to cause the network to output incorrect controls that cause the vehicle to deviate from its intended trajectory and crash.

3. Traffic light recognition

When traffic light recognition systems are deployed in new cities, fine-tuning is necessary for autonomous vehicles to adapt effectively. Patel et al. (2020) proposed a physical poisoning attack in the fine-tuning progress. They put electronic billboards near the traffic lights. They used the bait stage to launch when autonomous vehicles were being trained and then switched the stage to attack when the victim models were deployed. Due to the complexity of conducting physical experiments, this approach was executed within CARLA.

4. Trajectory prediction

GPS signals may be hindered by a high density of buildings in the urban area (Kos et al., 2010). In that case, the autonomous vehicle systems rely on image-based localization for self-position. Brand et al. (2022) designed an adversarial patch to attack the image-based localization module. According to the pasting of the adversarial patch on a street billboard in front of a traffic intersection, the vehicle's navigation system was disrupted. Zhang QZ et al. (2022) proposed the first adversarial attack that perturbs normal vehicle trajectories considering real-world constraints and impacts. They reported their attack performance on various trajectory prediction models (Rudenko et al., 2020) and datasets (Rasouli et al., 2019). Moreover, their attack method could achieve transfer-based attacks and query-based attacks at the same time. They also explored the potential methods against adversarial examples via data augmentation and trajectory smoothing.

Instead of generating adversarial point cloud examples, Li YM et al. (2021) proposed a trajectory attack named fool LiDAR perception with adversarial trajectory (FLAT), and the illustration of FLAT is shown in Fig. 11. They spoofed a self-driving car's trajectory with small perturbations, which is enough to make safety-critical objects undetectable or be detected with incorrect positions. Although FLAT uses LiDAR point clouds to attack the vehicle detection models, it can also affect various downstream tasks, i.e., 3D object detection (Pan et al., 2021), semantic segmentation (Zhang H et al., 2018), motion prediction (Fang et al., 2020), and multi-object tracking (Zeng et al., 2022).

3.2 LiDAR-based attacks

Cao et al. (2019) studied the adversarial vulnerability of the LiDAR-based perception system for ADSs. They constructed emergency brake attacks

and AV freezing attacks. Instead of painting a naturalistic pattern on the car to fool the camera sensors, Tu et al. (2020) decided to generate a 3D mesh and place it on the rooftop of a target vehicle. Their method can produce the 3D mesh in both white-box and black-box settings. The shape of a 3D mesh can also be chosen from common objects (couch, chair, table, bike, canoe, etc.). The visualization of the 3D mesh is shown in Fig. 12. Based on the observation that current LiDAR-based perception models do not learn occlusion information in the LiDAR point clouds, Sun et al. (2020) proposed a black-box adversarial attack to fool LiDAR sensors.

Yang KC et al. (2021) focused on the vulnerability of LiDAR sensors in ADSs and proposed an

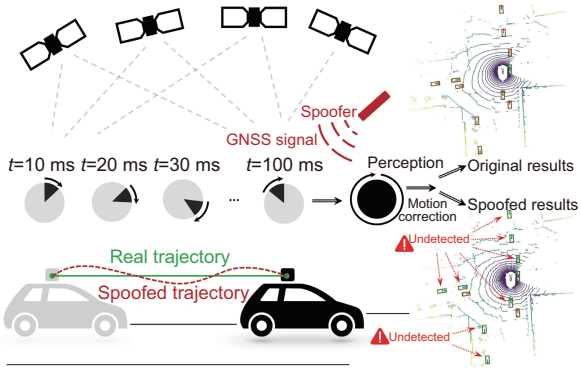

Fig. 11 Illustration of fool LiDAR perception with adversarial trajectory (FLAT) and the motion correction process (Li YM et al., 2021). The top right and bottom right figures are respectively the original and distorted LiDAR sweep and the detection results. Green/Red boxes denote the ground truth/prediction. In this example, FLAT makes the detector miss 8 out of 11 vehicles. GNSS: global navigation satellite system. References to color refer to the online version of this figure

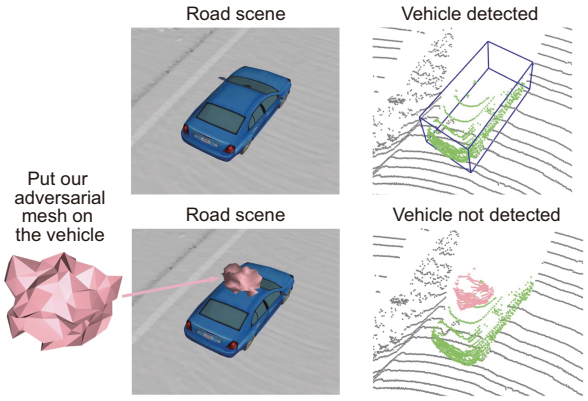

Fig. 12 A stop sign image was injected into a camera by a projector, which was detected by YOLOv3 (Tu et al., 2020)

adversarial attack in both white-box and black-box scenarios. They investigated the physical properties of 3D adversarial objects and the existence of adversarial points. They found that a small number of points from distant vehicles and occluded vehicles, which are normal, are also labeled as vehicles. They inferred that the adversary points generated from the objects can mimic these features so they can be misdetected as vehicles by the deep learning models, which may inspire the following defense strategies.

3.3 Multifusion attacks

Hallyburton et al. (2022) designed the frustum attack, which compromises camera–LiDAR fusion by exploiting the semantic consistencies between camera and LiDAR data, making it difficult for existing anti-spoofing defenses to detect and defend against this attack. Although the frustum attack achieved a satisfactory result against the camera and LiDAR sensors simultaneously, the feasibility needs more investigation when relative distance and angle change between the spoofer and the victim. Guesmi and Alouani (2022) investigated the adversarial vulnerability of ultra-wide band (UWB) radars on ADSs. They designed an adversarial radio noise attack (a-RNA) on radar-based environment perception systems. The proposed a-RNA attack involves injecting adversarial radio noise into the wireless channel to cause an obstacle recognition failure in UWB-based systems. They tested a-RNA on real-world UWB-based environment perception systems.

4 Digital-world scenario

In this section, we mainly discuss the adversarial attacks against ADSs in the digital-world scenario. These adversarial attacks are designed against various perception tasks in an ADS. However, due to the gap between the digital world and the physical world, these digital-world attack methods are less likely to attack the ADS successfully. We also investigate those attacks because they can inspire follow-up research on the adversarial vulnerability of ADSs.

4.1 Camera-based attacks

1. Trajectory prediction

Besides hijacking the attack choosing traffic lanes as adversarial examples to attack the trajectory

prediction of autonomous systems, AdvDO (Cao et al., 2022) manipulates the trajectory of the adversary agent to affect the Ego agent's trajectory. They devised an optimization-based adversarial attack framework that leverages a carefully designed differentiable dynamic model to generate realistic adversarial trajectories.

2. Vehicle detection

Zhang Y et al. (2019) designed a CAMOU attack to paint a unique pattern on the vehicle's body and hide it from being detected by surveillance cameras. Besides the main contribution of its first proposed vehicle detection attack, the main challenge is generating a complex pattern with the photorealistic Unreal-4 game engine. They arranged an approximate gradient network and repeatedly queried the target model to render adversarial camouflage with the Unreal-4 engine.

Following the prior research, Wang JK et al. (2021) claimed that the previous adversarial vehicle patterns do not exploit intrinsic characteristics, such as model-agnostic and human-specific patterns. These authors evaded the human-specific bottom–up attention and model-agnostic attention and proposed the dual attention suppression (DAS) attack to generate visually natural camouflages. The whole process is shown in Fig. 13.

Due to the unsatisfactory adversarial transferability of DAS, Wang DH et al. (2022) designed a robust full-coverage camouflage attack (FCA) to fool vehicle detectors. The experimental results illustrated that the transferability between different vehicle detectors was enhanced. Wu T et al. (2020b) proposed an enlarge-and-repeat (ER) attack to generate mosaic-like adversarial vehicle textures and applied it to the surface of a vehicle in the CARLA simulator. They used two main techniques, namely, an ER process and a discrete searching method, to craft these adversarial patterns. After querying the target model repeatedly, the mosaic-like adversarial patterns on vehicles can pose significant threats. However, the query-based adversarial strategy in ER has limitations due to the challenge of deploying it in real-world scenarios.

4.2 Multifusion attacks

Tu et al. (2022) used a differentiable method to render 3D adversarial mesh into both LiDAR and image inputs. The potential reason for such attacks may be the mapping between a mesh pixel and a LiDAR point far away from the mesh. Specifically, these false associations can easily occur if the assigned pixel for each LiDAR point is shifted by a few pixels, since objects that are far apart in a 3D space may appear very close in 2D images. The multimodal perception systems used in ADSs can be divided into two broad types: cascaded models, which use each modality independently, and fusion models, which learn from different modalities simultaneously. Abdelfattah et al. (2021a) placed a 3D adversarial mesh with learned shape and texture on top of a car (Abdelfattah et al., 2021b). The adversarial texture fools the camera detector, and the adversarial shape fools the LiDAR detector. Their

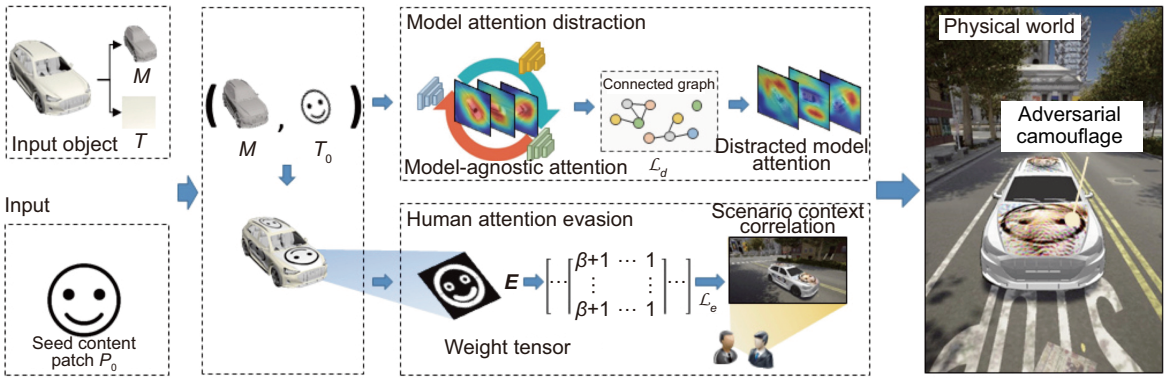

Fig. 13 The framework of the DAS method (Wang JK et al., 2021). Wang JK et al. (2021) first distracted the intrinsic attention characteristic by fully exploiting the similar attention patterns of models and forcing the "heat" regions away from the target object with loss function $\mathcal{L}_d$. Then they evaded the human-specific visual attention mechanism by correlating the appearance of adversaries to the context scenario and preserving the shape information of seed content image to generate visually-natural adversarial camouflage

3D adversarial mesh shows satisfactory performance on cascade models and needs more investigation on the fusion models.

Zhu ZJ et al. (2023) evaluated the adversarial vulnerability of autonomous systems, specifically in the context of 3D object detection with bird's-eye-view (BEV) representations. They observed that BEV models are more vulnerable to adversarial noise, such as ℓ_p adversarial perturbations and adversarial patches, compared to non-BEV models. The difference in adversarial robustness comes from the stages after feature extraction. Fusion models that combine both camera and LiDAR inputs generally outperform camera-only methods in terms of both performance and robustness. However, when perturbations are applied to both images and point clouds, BEVFusion, a fusion model using BEV features, becomes more vulnerable to adversarial attacks. Based on these observations, they designed a 3D consistent patch attack. This attack involves attaching adversarial patches in a way that ensures spatiotemporal consistency, making it more appropriate for autonomous driving scenarios. They generated adversarial examples by pasting the adversarial patches on objects in the real 3D space and projecting them into 2D images.

5 Robustness evaluations and defense strategies

In this section, we review the literature on ADSs in the context of physical adversarial robustness and the corresponding defense strategies. We analyze the research from two aspects, robustness evaluation and physical adversarial defenses. Specifically, in Section 5.1, we demonstrate the adversarial vulnerability of each sensor under ADSs. Thereafter, we discuss the adversarial defense strategies in ADSs in Section 5.2 considering three steps: input image preprocessing, adversarial example detection, and model enhancement.

5.1 Robustness evaluations

Recent research has demonstrated that ADSs are vulnerable to adversarial attacks in different scenarios. Based on the adversarial robustness of ADSs in various perception tasks, it is necessary to propose a comprehensive study on the adversarial robustness of ADSs and find potential ways to enhance it.

We investigate two primary types of physical adversarial attacks: perturbation attacks and patch attacks. Zhang JD et al. (2021) proposed an end-to-end evaluation framework to assess driving safety using a set of driving safety performance metrics. They observed that the deep stereo geometry network (DSGN) model (Chen YL et al., 2020) demonstrates stronger robustness to patch attacks compared to the stereo recurrent-convolutional neural network (R-CNN) model (Li PL et al., 2019). The potential reason for the stronger robustness of the DSGN model is the spatial pyramid pooling (SPP) module, which may restrain the impact of the adversarial patch on 3D object detection, hence enhancing its robustness to patch attacks. Rana and Madaan (2020) investigated the adversarial vulnerability on license plate recognition (LPR) models. They offered adversarial examples for which no defense mechanism currently exists.

Xu CJ et al. (2022) designed the SafeBench, the first unified platform for the safety evaluation of autonomous vehicles. SafeBench allows for the evaluation of AD algorithms on 2352 generated safety-critical testing scenarios, 4 scenario generation algorithms, and 10 diverse driving routes. The platform reports evaluation results based on 10 metrics, ranging from collision rate to route completion percentage. The authors believed that SafeBench would motivate the development of new testing scenario generation and safe AD algorithms. In the platform, they observed that MSF contributes to the robustness of object detection in autonomous driving algorithms.

The adversarial patch stands out as a well-known method of attack within the domain of physical adversarial challenges. Hingun et al. (2023) introduced the realistic adversarial patch (REAP) benchmark to facilitate the evaluation of such patch attacks. Leveraging the Mapillary Vistas Dataset (Ertler et al., 2020), REAP encompasses over 14 000 traffic signs, each meticulously modified with geometric and lighting transformations to enable the realistic application of digital patches. Insights gained from this benchmark reveal that when realistic constraints are applied, the efficacy of patch attacks is greatly diminished. Furthermore, it was found that standard adversarial training methods serve as a robust defense mechanism, mitigating the threat posed by the current patch attacks in practical settings.

5.2 Defense strategies

In this subsection, we discuss the model defense strategies in three categories: input image preprocessing, adversarial example detection, and model enhancement. The input image preprocessing strategies include preprocessing the input image to eliminate the adversarial features with different methods, i.e., completing and smoothing. The defense methods do not require retraining of the model, so that most of them could be incorporated into most existing autonomous driving perception systems. The adversarial detection method means that ADSs could detect the adversarial examples as abnormal. The model enhancement method needs to retrain the model, and it can recognize the adversarial examples as correct output. A summary of the referred papers on physical adversarial defenses is listed in Table 2. The table shows that the existing physical adversarial defenses are proposed mainly in model enhancement. The reason is that a few certified defense methods are designed against adversarial patches, which are currently the main form of physical adversarial attacks in ADSs.

1. Input image preprocessing

Hayes (2018) designed the digital watermarking (DW) defense by using the magnitude of the saliency maps to detect adversarial perturbations and remove them. Liu et al. (2022) proposed segment and complete (SAC) defense against the traffic sign adversarial examples and other physical patch based attacks. The key to SAC is to detect adversarial patches robustly. They trained a patch segmentor to segment adversarial patches from the inputs and generated a "completed patch mask," which is guaranteed to cover the entire adversarial patch. After masking the entire adversarial patches, the object detector could output the correct labels. They also presented the APRICOT-Mask dataset, which augments the APRICOT dataset with pixel-level annotations of adversarial patches. Xu H et al. (2021) used a lane verification model, which applies to any existing lane detection models, to defend against adversarial attacks in lane detection for ADSs. The details of their framework are illustrated in Fig. 14.

Recent studies on adversarial weather attacks (Marchisio et al., 2022; Schmalfuss et al., 2023) have demonstrated the capability of generating adversarial examples that mimic rain-like conditions.

Table 2 Physical adversarial defense methods mentioned in this paper

Defense	Sensor		Defense strategy		
	Camera	LiDAR	Model enhancement	Input image preprocessing	Adversarial example detection
SAC (Liu et al., 2022)	✓			✓	
Lagraa et al. (2019)'s	✓				✓
SVF (Sun et al., 2020)		✓	✓		
CARLO (Sun et al., 2020)		✓			✓
Demasked smoothing (Yatsura et al., 2023)	✓		✓		
T-GP (Han et al., 2021)	✓	✓			✓
PatchCensor (Huang YH et al., 2023)	✓		✓		
ECViT (Chen ZY et al., 2022)	✓		✓		
DeRS (Levine and Feizi, 2020)	✓		✓		
IBP to patch (Chiang et al., 2020)	✓		✓		
BAGCERT (Metzen and Yatsura, 2021)	✓		✓		
BlurNet (Raju and Lipasti, 2020)	✓		✓		
Xu H et al. (2021)'s	✓			✓	
DW (Hayes, 2018)	✓			✓	
SentiNet (Chou et al., 2020)	✓				✓
DOA (Wu T et al., 2020a)	✓		✓		
Ad-YOLO (Ji et al., 2021)	✓				✓
MAT (Metzen et al., 2021)	✓		✓		
Hau et al. (2022)'s	✓				✓
Yu et al. (2022)'s	✓		✓		
FPDA (Rossolini et al., 2023)	✓				✓
iGAT (Deng and Mu, 2023)	✓		✓		
Jin et al. (2022)'s	✓				✓

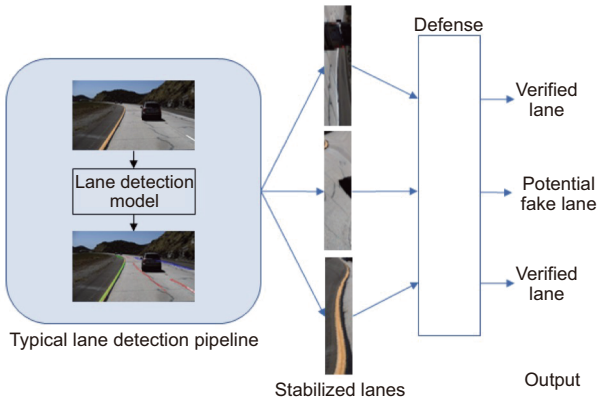

Fig. 14 The defense framework in Xu H et al. (2021). The framework first stabilizes lanes in the typical lane detection pipeline. Then, it uses a lane verification model to detect potential fake lanes

The "fakeWeather" research underscores that conventional rain-removal strategies are insufficient in countering these adversarial attacks. In response, Yu et al. (2022) conducted a comprehensive analysis of the modules used in existing deraining methods. Through this analysis, they identified key modules that substantially enhance the robustness of deraining techniques. Building on these findings, they developed a refined deraining model that incorporates these effective modules. The efficacy of this new model was then rigorously evaluated against contemporary adversarial examples simulating rainy conditions.

2. Adversarial example detection

Lagraa et al. (2019) proposed a method for detecting adversarial examples in a robotic operating system (ROS) embedded in a self-driving car. Their method can attack various types of attacks. However, they assumed that the ROS camera node was under attack and that the attacker was in the middle between the camera node and the processing node. The abnormal adversarial examples they found are still generated by digital attacks. Their performances under real-world scenarios need more investigation. Sun et al. (2020) observed that current LiDAR-based perception models do not learn occlusion information in the LiDAR point clouds. Therefore, they proposed the CARLO (short for oCclusion-Aware hieRarchy anomaLy detectiOn) method as a model-agnostic plug-in to detect adversarial LiDAR point clouds. CARLO consists of two primary components: free space detection (FSD) and laser penetration detection (LPD). FSD and LPD are designed to exploit occlusion-related characteristics: the free space inside a detected bounding box and the locations of points inside the frustum corresponding to a detected bounding box. CARLO treats ignored occlusion patterns as invariant physical features and uses them to detect spoofed data.

Han et al. (2021) proposed a one-class classification model based on the Transformer for anomaly detection. The detection methodology, named Transformer with gradient penalty (T-GP), can successfully detect the localization attack, the lane detection attack, and the traffic sign attacks at the same time. Although T-GP could detect MSF attacks in an LG Silicon Valley Lab (LGSVL) simulator, the adversarial examples that they generated are still based on the pixel-to-pixel digital strategy. It is a challenge to evaluate their performance against stronger physical-based adversarial examples. Chou et al. (2020) observed the threats from adversarial patches and designed SentiNet to detect those contiguous portions. SentiNet uses gradient-weighted class activation mapping (Grad-CAM) to obtain the mask. The runtime of SentiNet is about 2.5 s, which may be acceptable for other tasks. However, the latency of SentiNet needs to be sped up to satisfy the autonomous driving perception tasks.

Ji et al. (2021) built a plug-in defense component on top of the YOLOv2 detection system, named Ad-YOLO, which helps human detectors against patch attacks. Ad-YOLO adds a patch class to the YOLO architecture, allowing it to detect both objects of interest and adversarial patches. The main idea is to enhance the resilience of the YOLOv2 model against patch attacks by recognizing adversarial patches as a new category and restoring the original object information. The whole structure is illustrated in Fig. 15.

Hau et al. (2022) observed that object-hiding attacks still induce shadow artifacts in the 3D point clouds. Therefore, they used LiDAR sensors to enhance the adversarial robustness of cameras. By using 3D shadows as a physical invariant, the proposed methodology can detect objects that are hidden by adversaries in the scene. Jin et al. (2022) found out that the motion of each surrounding object is affected by its neighbors and the overall traffic scenario information. This means that the behavior of each object is influenced by the presence and movements of other nearby objects, as well as the larger context

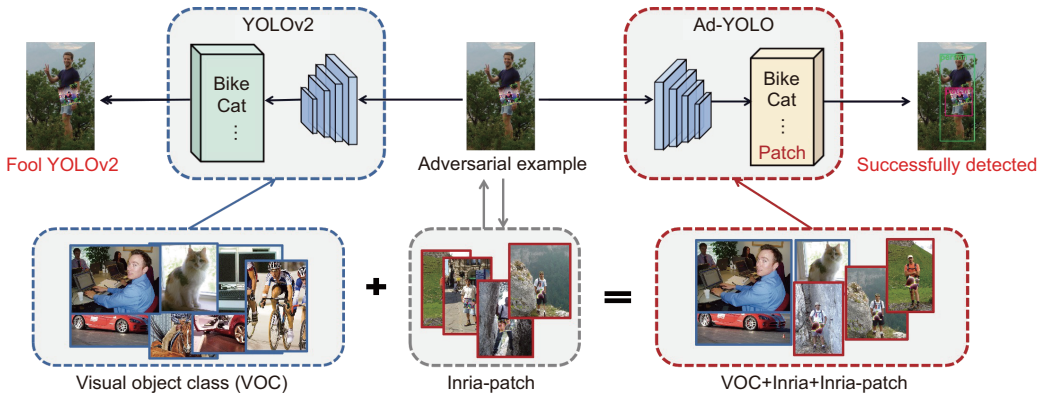

Fig. 15 Typical adversarial example detection model structure (Ad-YOLO) (Ji et al., 2021). The aim of this structure is to identify the adversarial example. One common method is to add one "adversarial example" category to the last layer of the object detector

of the traffic scenario. Inspired by this observation, they designed a graph neural network (GNN) based relation learning network to detect abnormal perception information in self-driving scenarios. The GNN helps build relationships and aggregate multimodal information to improve the robustness and safety of self-driving systems.

Rossolini et al. (2023) developed the fast patch detection algorithm (FPDA) to identify semantic segmentation predictions affected by adversarial patches. This algorithm specifically targets the identification of overactivated internal features, which are characteristic markers of adversarial attacks. Using a methodology that integrates feature compression, normalization, and threshold-based filtering, the FPDA was engineered to facilitate the real-time detection of adversarial patches. To substantiate the efficacy and real-time applicability of the FPDA, they conducted comprehensive experimental evaluations. These evaluations not only confirm the effectiveness of the FPDA but also include a detailed comparative analysis with the foundational hyperneuron method. Furthermore, the research explores the FPDA's performance in scenarios involving defense-aware attacks, thereby providing a holistic assessment of its practical utility in adversarial contexts.

3. Model enhancement

Wu T et al. (2020a) first proposed a defense method named defense against occlusion attack (DOA) to defend image classifiers against physical adversarial traffic signs. DOA uses adversarial strategies to stick adversarial patches on benign examples and uses adversarial training to enhance the robustness of the image classification model. Sun et al. (2020) designed sequential view fusion (SVF) as a robust LiDAR-based perception architecture for ADSs. It addresses the vulnerability of current LiDAR-based perception systems by sequentially fusing the front view (FV) and 3D representations of LiDAR point clouds during the end-to-end learning process. By ensuring that the features from the FV are adequately used, SVF reduces the mean success rate of attacks significantly.

Metzen et al. (2021) designed meta-adversarial training (MAT) to enhance the adversarial robustness of autonomous driving perception systems. This method randomly assigns a target class and a fixed step size to each meta-patch in the set of meta-patches. The random assignment of targets and step sizes promotes diversity among the meta-patches, ensuring a wide range of patch patterns and preventing overfitting to a single meta-patch. The illustration is shown in Fig. 16. Choi and Tian (2022) proposed a new objectness-aware adversarial training approach for YOLO detectors to address the vulnerability under autonomous driving scenarios. The proposed defense approach showcases improved robustness against objectness-oriented attacks.

After Cao et al. (2023) proposed a physical removal attack (PRA) method to attack LiDAR sensors, they extended the methodology of Hau et al. (2022) to include PRA detection, named fake shadow detection (FSD). The FSD methodology first consists of identifying the point cloud's shadow regions in the region of interest (ROI) to find potential object removal attacks (ORAs). They also proposed

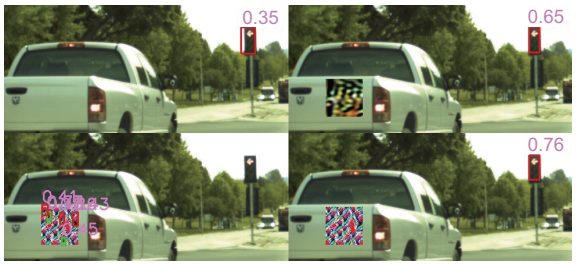

Fig. 16 Illustration of patch attack against the undefended model (left) and the defended model with meta-adversarial training (right) (Metzen et al., 2021)

azimuth-based detection to inspect the horizontal angular view of the LiDAR (azimuth).

The ensemble-based defense strategy (Croce and Hein, 2020) involves training multiple base classifiers to work together in defending against adversarial attacks. This cooperative approach aims to improve the statistical stability and cooperation between the base models, ultimately enhancing the robustness of the ensemble against adversarial attacks. The interactive global adversarial training (iGAT) (Deng and Mu, 2023) method includes a probabilistic distribution rule for selectively allocating global adversarial examples to train base classifiers and a regularization penalty for addressing vulnerabilities across all base classifiers. While the iGAT strategy uses an ensemble-based approach to improve adversarial robustness, its increased time consumption compared to other model enhancement methods renders it less effective for large-scale autonomous perception models.

The physical adversarial attacks usually generate patch-based perturbations and paste them on traffic elements. The problem has attracted significant interest in the research community since vision Transformers (ViTs) (Dosovitskiy et al., 2020) have been used to solve the perception problem in ADSs. Recently, a few researchers have proposed certified defense methods against patch-based attacks, and they claimed that their method can inspire the defense in ADSs (Raju and Lipasti, 2020; Metzen and Yatsura, 2021).

Chiang et al. (2020) extended interval bound propagation (IBP) defense (Gowal et al., 2019) against patch attacks. Gu et al. (2022) pointed out that ViT is significantly more vulnerable against physical domain adversarial patches. Huang YH et al. (2023) proposed PatchCensor to certify the patch robustness of ViT by applying exhaustive testing. Chen ZY et al. (2022) proposed an efficient certifiable ViT (ECViT) to capture more discriminable local context of an image while preserving the global semantic information, which leads to better robustness. They proved by experiments that PatchCensor and ECViT could defend against physical adversarial patches. Levine and Feizi (2020) designed (de)randomized smoothing to de-randomized patch adversarial attacks and guaranteed that no patch adversarial examples exist for the classification model. In this way, the number of smoothed images affected by the adversarial patch can be calculated. This quantitative framework is beneficial for the improvement of model robustness. Yatsura et al. (2023) proposed demasked smoothing against adversarial patches on the semantic segmentation module.

6 Future discussion

We present some of the directions for future consideration. These future directions discuss the physical adversarial vulnerability of ADSs considering experimental settings, adversarial attacks, and adversarial defenses. The details are shown below:

1. The adversarial vulnerability in ADSs lacks united platforms and comprehensive benchmarks. Currently, ADS evaluations vary widely, using diverse models, datasets, and scenarios. This variation prevents standardized comparisons across studies, undermining efforts to assess and improve ADS security holistically. The lack of integrated evaluation frameworks hampers the understanding and mitigation of adversarial threats effectively. Establishing unified benchmarks and platforms is critical for reliable comparisons and enhancements in ADS safety. Future research must prioritize these integrated systems to safeguard autonomous vehicles against evolving adversarial challenges, ensuring their reliability and security in complex environments.

2. The adversarial transferability of physical adversarial attacks under ADS requires more investigation. Research has shown that the effectiveness of physical adversarial attacks often diminishes when applied across different models, especially if the surrogate model used in the attack's design significantly differs from the target model. This limitation suggests that current methods for creating adversarial examples are too specific to certain model architectures, reducing their effectiveness universally. Future

studies should focus on improving the transferability of physical adversarial attacks across models by developing methods that exploit common vulnerabilities rather than relying on model-specific weaknesses. Enhancing this aspect of adversarial robustness is crucial for securing ADSs against a wider array of threats and ensuring their safe operation across various environments.

3. The deployment and evaluation of query-based attacks in real-world scenarios present significant practical challenges. These attacks, effective in controlled simulations and digital environments, face obstacles under real-world conditions due to environmental unpredictability, including varying lighting, weather, and physical obstructions. Additionally, implementing these attacks discreetly while navigating physical access limitations complicates their integration. This discrepancy highlights the difficulty in optimizing adversarial perturbations outside of simulations, pointing to a gap between theoretical effectiveness and real-world applicability. It underscores the urgency in developing strategies that consider the complexity of real-world scenarios, aiming to bolster the robustness and reliability of ADS against sophisticated adversarial techniques.

4. The cost and technical difficulty of evaluating ADS perception tasks in the real world are high, except for traffic sign recognition and person detection tasks. The unpredictability and diversity of real-world scenarios demand comprehensive datasets for accurate testing, while the integration of advanced sensor technologies like LiDAR, radar, and cameras requires complex testing setups. Additionally, safety and logistical issues of real-world testing increase complexity and expense. There is a crucial need for research on simple, cost-effective attack strategies that can be realistically implemented and assessed. Developing such strategies could lower barriers, offer realistic robustness evaluations, and enhance the safety and security of ADS technologies.

5. The stealthiness of physical adversarial examples in ADS needs to be enhanced. Current attempts to make these attacks unnoticeable often fail, drawing unwanted human attention and compromising their undetectability. The challenge lies in making adversarial examples blend seamlessly into the driving environment without alerting drivers. Future research needs to focus on advanced methods that evade detection by both ADS sensors and human observers. This involves careful manipulation of characteristics such as color, shape, size, and placement to reduce visibility. Using machine learning and image processing to predict and counter human responses to these changes is vital. Enhancing stealthiness will bolster ADS security, maintaining trust and safety in these technologies.

6. ADSs lack effective and common defense strategies, with most defenses tested in simulations rather than real-world settings. This oversight is critical, considering the variety of perception modules in ADSs, each vulnerable to different adversarial attacks. Future research should aim to understand how adversarial examples affect various ADS sensors and develop comprehensive defense strategies that ensure protection across all sensors and scenarios. This necessitates a detailed evaluation of current defenses to pinpoint their strengths and weaknesses, followed by innovation to improve their real-world applicability. Establishing common defense mechanisms will enhance ADS resilience, ensuring their safety and reliability against evolving adversarial threats in diverse conditions.

7. The adversarial vulnerabilities of radar sensors are notably under-explored. Radars are crucial in ADSs for providing essential data on distance and radial velocity, crucial for vehicle orientation predictions. High-resolution radars can even discern pedestrian gestures and directions by analyzing body part velocities. These detailed data are vital for the safety and efficiency of ADSs. Despite their importance, the susceptibility of radar sensors to adversarial attacks remains largely unexplored. Future research needs to prioritize understanding these vulnerabilities to develop stronger defense mechanisms, ensuring ADSs' reliability and safety in various conditions.

8. Future research needs to investigate plugin-and-play strategies to help ADS defend against physical adversarial examples. Traditional digital defenses require significant computational resources, a demand that grows with model complexity and dataset size, becoming unsustainable. Plugin-and-play approaches promise effective, resource-efficient defenses that easily integrate with existing ADS frameworks, facilitating rapid adaptation to new threats. These strategies must be lightweight and be easily updated, focusing on ease of use, scalability, and low computational demands. Advancing plugin-and-play defenses is vital for ADS security, enabling

robust protection against evolving adversarial tactics without excessive resource use, thus ensuring the safety and integrity of autonomous vehicles in dynamic adversarial landscapes.

7 Conclusions

In this survey, we present a comprehensive analysis of the physical adversarial attacks and defenses in the ADS area. We categorize the physical adversarial attacks considering attack scenarios, sensors, and tasks. Subsequently, we divide the physical adversarial defenses in terms of input image preprocessing, adversarial example detection, and model enhancement. We discuss the current limits and future directions about the adversarial vulnerability in ADSs. We hope this survey can bring inspiration to the autonomous driving community for future directions.

Contributors

Shuai ZHAO and Boyuan ZHANG designed the framework and drafted the paper. Yucheng SHI helped organize the paper. Yang ZHAI collected the literature and optimized the layout of pictures. Yahong HAN and Qinghua HU revised and finalized the paper.

Conflict of interest

Yahong HAN is a corresponding expert of *Frontiers of Information Technology & Electronic Engineering*, and he was not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

References

Abdelfattah M, Yuan KW, Wang ZJ, et al., 2021a. Adversarial attacks on camera-LiDAR models for 3D car detection. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.2189-2194. https://doi.org/10.1109/IROS51168.2021.9636638

Abdelfattah M, Yuan KW, Wang ZJ, et al., 2021b. Towards universal physical attacks on cascaded camera-LiDAR 3D object detection models. IEEE Int Conf on Image Processing, p.3592-3596. https://doi.org/10.1109/ICIP42928.2021.9506016

Ansari MA, Singh DK, 2021. Human detection techniques for real time surveillance: a comprehensive survey. *Multim Tools Appl*, 80(6):8759-8808. https://doi.org/10.1007/s11042-020-10103-4

Bai T, Luo JQ, Zhao J, 2022. Inconspicuous adversarial patches for fooling image-recognition systems on mobile devices. *IEEE Int Things J*, 9(12):9515-9524. https://doi.org/10.1109/JIOT.2021.3124815

Benz P, Zhang CN, Imtiaz T, et al., 2020. Double targeted universal adversarial perturbations. Proc 15th Asian Conf on Computer Vision, p.284-300. https://doi.org/10.1007/978-3-030-69538-5_18

Boloor A, He X, Gill C, et al., 2019. Simple physical adversarial examples against end-to-end autonomous driving models. IEEE Int Conf on Embedded Software and Systems, p.1-7. https://doi.org/10.1109/ICESS.2019.8782514

Boloor A, Garimella K, He X, et al., 2020. Attacking vision-based perception in end-to-end autonomous driving models. *J Syst Archit*, 110:101766. https://doi.org/10.1016/j.sysarc.2020.101766

Brand M, Naeh I, Teitelman D, 2022. Adversarial attack against image-based localization neural networks. https://arxiv.org/abs/2210.06589

Brock A, Donahue J, Simonyan K, 2018. Large scale GAN training for high fidelity natural image synthesis. 7th Int Conf on Learning Representations.

Cao YL, Xiao CW, Cyr B, et al., 2019. Adversarial sensor attack on LiDAR-based perception in autonomous driving. Proc ACM SIGSAC Conf on Computer and Communications Security, p.2267-2281. https://doi.org/10.1145/3319535.3339815

Cao YL, Wang NF, Xiao CW, et al., 2021. Invisible for both camera and LiDAR: security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. IEEE Symp on Security and Privacy, p.176-194. https://doi.org/10.1109/SP40001.2021.00076

Cao YL, Xiao CW, Anandkumar A, et al., 2022. AdvDO: realistic adversarial attacks for trajectory prediction. 17th European Conf on Computer Vision, p.36-52. https://doi.org/10.1007/978-3-031-20065-6_3

Cao YL, Bhupathiraju SH, Naghavi P, et al., 2023. You can't see me: physical removal attacks on LiDAR-based autonomous vehicles driving frameworks. 32nd USENIX Security Symp, p.2993-3010.

Chaubey A, Agrawal N, Barnwal K, et al., 2020. Universal adversarial perturbations: a survey. https://arxiv.org/abs/2005.08087

Chen J, Gao Y, Liu Y, et al., 2022. Leveraging model poisoning attacks on license plate recognition systems. IEEE Int Conf on Trust, Security and Privacy in Computing and Communications, p.827-834. https://doi.org/10.1109/TrustCom56396.2022.00115

Chen ST, Cornelius C, Martin J, et al., 2019. ShapeShifter: robust physical adversarial attack on faster R-CNN object detector. European Conf on Machine Learning and Knowledge Discovery in Databases, p.52-68. https://doi.org/10.1007/978-3-030-10925-7_4

Chen YL, Liu S, Shen XY, et al., 2020. DSGN: deep stereo geometry network for 3D object detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.12536-12545. https://doi.org/10.1109/CVPR42600.2020.01255

Chen ZY, Li B, Xu JH, et al., 2022. Towards practical certifiable patch defense with vision transformer. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15127-15137. https://doi.org/10.1109/CVPR52688.2022.01472

Cheng ZY, Liang J, Choi H, et al., 2022. Physical attack on monocular depth estimation with optimal adversarial patches. 17th European Conf on Computer Vision,

p.514-532. https://doi.org/10.1007/978-3-031-19839-7_30

Chiang PY, Ni RK, Abdelkader A, et al., 2020. Certified defenses for adversarial patches. 8th Int Conf on Learning Representations.

Choi JI, Tian Q, 2022. Adversarial attack and defense of YOLO detectors in autonomous driving scenarios. IEEE Intelligent Vehicles Symp, p.1011-1017. https://doi.org/10.1109/IV51971.2022.9827222

Chou E, Tramèr F, Pellegrino G, 2020. SentiNet: detecting localized universal attacks against deep learning systems. IEEE Security and Privacy Workshops, p.48-54. https://doi.org/10.1109/SPW50608.2020.00025

Croce F, Hein M, 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. Proc 37th Int Conf on Machine Learning, p.2206-2216.

Deng Y, Mu TT, 2023. Understanding and improving ensemble adversarial defense. 37th Int Conf on Neural Information Processing Systems.

Dibaei M, Zheng X, Jiang K, et al., 2020. Attacks and defences on intelligent connected vehicles: a survey. *Digit Commun Netw*, 6(4):399-421. https://doi.org/10.1016/j.dcan.2020.04.007

Ding WH, Xu CJ, Arief M, et al., 2023. A survey on safety-critical driving scenario generation—a methodological perspective. *IEEE Trans Intell Transp Syst*, 24(7):6971-6988. https://doi.org/10.1109/TITS.2023.3259322

Dosovitskiy A, Beyer L, Kolesnikov A, et al., 2020. An image is worth 16×16 words: transformers for image recognition at scale. 9th Int Conf on Learning Representations.

Duan RJ, Ma XJ, Wang YS, et al., 2020. Adversarial camouflage: hiding physical-world attacks with natural styles. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.997-1005. https://doi.org/10.1109/CVPR42600.2020.00108

Duan RJ, Mao XF, Qin AK, et al., 2021. Adversarial laser beam: effective physical-world attack to DNNs in a blink. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.16057-16066. https://doi.org/10.1109/CVPR46437.2021.01580

Ertler C, Mislej J, Ollmann T, et al., 2020. The Mapillary Traffic Sign Dataset for detection and classification on a global scale. 16th European Conf on Computer Vision, p.68-84. https://doi.org/10.1007/978-3-030-58592-1_5

Eykholt K, Evtimov I, Fernandes E, et al., 2018a. Physical adversarial examples for object detectors. Proc 12th USENIX Workshop on Offensive Technologies, p.1. https://arxiv.org/pdf/1807.07769

Eykholt K, Evtimov I, Fernandes E, et al., 2018b. Robust physical-world attacks on deep learning visual classification. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1625-1634. https://doi.org/10.1109/CVPR.2018.00175

Fang LJ, Jiang QH, Shi JP, et al., 2020. TPNet: trajectory proposal network for motion prediction. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6796-6805. https://doi.org/10.1109/CVPR42600.2020.00683

Feng WW, Xu NQ, Zhang TZ, et al., 2024. Robust and generalized physical adversarial attacks via Meta-GAN. *IEEE Trans Inform Forensic Secur*, 19:1112-1125. https://doi.org/10.1109/TIFS.2023.3288426

Gowal S, Dvijotham K, Stanforth R, et al., 2019. Scalable verified training for provably robust image classification. Proc IEEE/CVF Int Conf on Computer Vision, p.4841-4850. https://doi.org/10.1109/ICCV.2019.00494

Gu JD, Tresp V, Qin Y, 2022. Are vision transformers robust to patch perturbations? 17th European Conf on Computer Vision, p.404-421. https://doi.org/10.1007/978-3-031-19775-8_24

Guesmi A, Alouani I, 2022. Adversarial attack on radar-based environment perception systems. https://doi.org/10.48550/arXiv.2211.01112

Hallyburton RS, Liu YP, Cao YL, et al., 2022. Security analysis of camera-LiDAR fusion against black-box attacks on autonomous vehicles. 31st USENIX Security Symp, p.1903-1920.

Han XS, Chen KJ, Zhou Y, et al., 2021. A unified anomaly detection methodology for lane-following of autonomous driving systems. IEEE Int Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking, p.836-844. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00119

Han XS, Xu GW, Zhou Y, et al., 2022. Physical backdoor attacks to lane detection systems in autonomous driving. Proc 30th ACM Int Conf on Multimedia, p.2957-2968. https://doi.org/10.1145/3503161.3548171

Hau Z, Demetriou S, Lupu EC, 2022. Using 3D shadows to detect object hiding attacks on autonomous vehicle perception. IEEE Security and Privacy Workshops, p.229-235. https://doi.org/10.1109/SPW54247.2022.9833890

Hayes J, 2018. On visible adversarial perturbations & digital watermarking. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops, p.1597-1604. https://doi.org/10.1109/CVPRW.2018.00210

Hingun N, Sitawarin C, Li J, et al., 2023. REAP: a large-scale realistic adversarial patch benchmark. Proc IEEE/CVF Int Conf on Computer Vision, p.4617-4628. https://doi.org/10.1109/ICCV51070.2023.00428

Hu YCT, Chen JC, Kung BH, et al., 2021. Naturalistic physical adversarial patch for object detectors. Proc IEEE/CVF Int Conf on Computer Vision, p.7828-7837. https://doi.org/10.1109/ICCV48922.2021.00775

Hu ZH, Huang SY, Zhu XP, et al., 2022. Adversarial texture for fooling person detectors in the physical world. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.13297-13306. https://doi.org/10.1109/CVPR52688.2022.01295

Huang LF, Gao CY, Zhou YY, et al., 2020. Universal physical camouflage attacks on object detectors. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.717-726. https://doi.org/10.1109/CVPR42600.2020.00080

Huang YH, Ma L, Li YC, 2023. PatchCensor: patch robustness certification for transformers via exhaustive testing. *ACM Trans Soft Eng Methodol*, 32(6):154. https://doi.org/10.1145/3591870

Jan STK, Messou J, Lin YC, et al., 2019. Connecting the digital and physical world: improving the robustness of adversarial attacks. Proc 33rd AAAI Conf on Artificial

Intelligence, p.962-969.
https://doi.org/10.1609/aaai.v33i01.3301962

Ji N, Feng YF, Xie HD, et al., 2021. Adversarial YOLO: defense human detection patch attacks via detecting adversarial patches. https://arxiv.org/abs/2103.08860

Jia W, Lu ZJ, Zhang HC, et al., 2022. Fooling the eyes of autonomous vehicles: robust physical adversarial examples against traffic sign recognition systems. https://arxiv.org/abs/2201.06192v1

Jin KF, Wang HY, Liu CX, et al., 2022. Graph neural network based relation learning for abnormal perception information detection in self-driving scenarios. Int Conf on Robotics and Automation, p.8943-8949. https://doi.org/10.1109/ICRA46639.2022.9812411

Kiran BR, Sobh I, Talpaert V, et al., 2022. Deep reinforcement learning for autonomous driving: a survey. *IEEE Trans Intell Transp Syst*, 23(6):4909-4926. https://doi.org/10.1109/TITS.2021.3054625

Kong ZL, Guo JF, Li A, et al., 2020. PhysGAN: generating physical-world-resilient adversarial examples for autonomous driving. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.14242-14251. https://doi.org/10.1109/CVPR42600.2020.01426

Kos T, Markezic I, Pokrajcic J, 2010. Effects of multipath reception on GPS positioning performance. Proc ELMAR, p.399-402.

Kumar KN, Vishnu C, Mitra R, et al., 2020. Black-box adversarial attacks in autonomous vehicle technology. IEEE Applied Imagery Pattern Recognition Workshop, p.1-7. https://doi.org/10.1109/AIPR50011.2020.9425267

Kwon H, Baek JW, 2021. Adv-Plate attack: adversarially perturbed plate for license plate recognition system. *J Sens*, 2021:6473833. https://doi.org/10.1155/2021/6473833

Lagraa S, Cailac M, Rivera S, et al., 2019. Real-time attack detection on robot cameras: a self-driving car application. 3rd IEEE Int Conf on Robotic Computing, p.102-109. https://doi.org/10.1109/IRC.2019.00023

Levine A, Feizi S, 2020. (De)Randomized smoothing for certifiable defense against patch attacks. Proc 34th Int Conf on Neural Information Processing Systems, Article 542.

Li PL, Chen XZ, Shen SJ, 2019. Stereo R-CNN based 3D object detection for autonomous driving. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7636-7644. https://doi.org/10.1109/CVPR.2019.00783

Li YM, Wen CC, Juefei-Xu F, et al., 2021. Fooling LiDAR perception via adversarial trajectory perturbation. Proc IEEE/CVF Int Conf on Computer Vision, p.7878-7887. https://doi.org/10.1109/ICCV48922.2021.00780

Liu J, Levine A, Lau CP, et al., 2022. Segment and complete: defending object detectors against adversarial patch attacks with robust patch detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.14953-14962. https://doi.org/10.1109/CVPR52688.2022.01455

Lovisotto G, Turner H, Sluganovic I, et al., 2021. SLAP: improving physical adversarial examples with short-lived adversarial perturbations. 30th USENIX Security Symp, p.1865-1882.

Lu JJ, Sibai H, Fabry E, 2017. Adversarial examples that fool detectors. https://arxiv.org/abs/1712.02494

Machado GR, Silva E, Goldschmidt RR, 2021. Adversarial machine learning in image classification: a survey toward the defender's perspective. *ACM Comput Surv*, 55(1):8. https://doi.org/10.1145/3485133

Man YM, Li M, Gerdes R, 2020. GhostImage: remote perception attacks against camera-based image classification systems. 23rd Int Symp on Research in Attacks, Intrusions and Defenses, p.317-332. https://arxiv.org/abs/2001.07792v2

Marchisio A, Caramia G, Martina M, et al., 2022. fakeWeather: adversarial attacks for deep neural networks emulating weather conditions on the camera lens of autonomous systems. Int Joint Conf on Neural Networks, p.1-9. https://doi.org/10.1109/IJCNN55064.2022.9892612

Metzen JH, Yatsura M, 2021. Efficient certified defenses against patch attacks on image classifiers. 9th Int Conf on Learning Representations.

Metzen JH, Finnie N, Hutmacher R, 2021. Meta adversarial training against universal patches. https://arxiv.org/abs/2101.11453

Modas A, Sanchez-Matilla R, Frossard P, et al., 2020. Toward robust sensing for autonomous vehicles: an adversarial perspective. *IEEE Signal Process Mag*, 37(4):14-23. https://doi.org/10.1109/MSP.2020.2985363

Moosavi-Dezfooli SM, Fawzi A, Fawzi O, et al., 2017. Universal adversarial perturbations. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.86-94. https://doi.org/10.1109/CVPR.2017.17

Pan XR, Xia ZF, Song SJ, et al., 2021. 3D object detection with Pointformer. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7459-7468. https://doi.org/10.1109/CVPR46437.2021.00738

Patel N, Krishnamurthy P, Garg S, et al., 2020. Bait and switch: online training data poisoning of autonomous driving systems. https://arxiv.org/abs/2011.04065

Qian YG, Ma DF, Wang B, et al., 2020. Spot evasion attacks: adversarial examples for license plate recognition systems with convolutional neural networks. *Comput Secur*, 95:101826. https://doi.org/10.1016/j.cose.2020.101826

Qiu SL, Liu QH, Zhou SJ, et al., 2019. Review of artificial intelligence adversarial attack and defense technologies. *Appl Sci*, 9(5):909. https://doi.org/10.3390/app9050909

Raju RS, Lipasti M, 2020. BlurNet: defense by filtering the feature maps. 50th Annual IEEE/IFIP Int Conf on Dependable Systems and Networks Workshops, p.38-46. https://doi.org/10.1109/DSN-W50199.2020.00016

Rana K, Madaan R, 2020. Evaluating effectiveness of adversarial examples on state of art license plate recognition models. IEEE Int Conf on Intelligence and Security Informatics, p.1-3. https://doi.org/10.1109/ISI49825.2020.9280477

Rasouli A, Kotseruba I, Kunic T, et al., 2019. PIE: a large-scale dataset and models for pedestrian intention estimation and trajectory prediction. Proc IEEE/CVF Int Conf on Computer Vision, p.6261-6270. https://doi.org/10.1109/ICCV.2019.00636

Rossolini G, Nesti F, D'Amico G, et al., 2023. On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving. *IEEE Trans Neur Netw Learn Syst*, 35(12):18328-18342. https://doi.org/10.1109/TNNLS.2023.3314512

Rudenko A, Palmieri L, Herman M, et al., 2020. Human motion trajectory prediction: a survey. *Int J Rob Res*, 39(8):895-935. https://doi.org/10.1177/0278364920917446

Sato T, Shen JJ, Wang NF, et al., 2021. Dirty road can attack: security of deep learning based automated lane centering under physical-world attack. Proc 30th USENIX Security Symp, p.3309-3326.

Schmalfuss J, Mehl L, Bruhn A, 2023. Distracting downpour: adversarial weather attacks for motion estimation. Proc IEEE/CVF Int Conf on Computer Vision, p.10072-10082. https://doi.org/10.1109/ICCV51070.2023.00927

Serban A, Poll E, Visser J, 2021. Adversarial examples on object recognition: a comprehensive survey. *ACM Comput Surv*, 53(3):66. https://doi.org/10.1145/3398394

Sitawarin C, Bhagoji AN, Mosenia A, et al., 2018. Rogue signs: deceiving traffic sign recognition with malicious ads and logos. https://arxiv.org/abs/1801.02780

Sun JS, Cao YL, Chen QA, et al., 2020. Towards robust LiDAR-based perception in autonomous driving: general black-box adversarial sensor attack and countermeasures. 29th USENIX Security Symp, p.877-894.

Suryanto N, Kim Y, Kang H, et al., 2022. DTA: physical camouflage attacks using differentiable transformation network. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15284-15293. https://doi.org/10.1109/CVPR52688.2022.01487

Szegedy C, Zaremba W, Sutskever I, et al., 2014. Intriguing properties of neural networks. https://doi.org/10.48550/arXiv.1312.6199

Thys S, Van Ranst W, Goedemé T, 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops, p.49-55. https://doi.org/10.1109/CVPRW.2019.00012

Tu J, Ren M, Manivasagam S, et al., 2020. Physically realizable adversarial examples for LiDAR object detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.13713-13722. https://doi.org/10.1109/CVPR42600.2020.01373

Tu J, Li HC, Yan XC, et al., 2022. Exploring adversarial robustness of multi-sensor perception systems in self driving. Proc 5th Conf on Robot Learning, p.1013-1024.

Wang DH, Jiang TS, Sun JL, et al., 2022. FCA: learning a 3D full-coverage vehicle camouflage for multi-view physical adversarial attack. Proc 36th AAAI Conf on Artificial Intelligence, p.2414-2422. https://doi.org/10.1609/aaai.v36i2.20141

Wang JK, Liu AS, Yin ZX, et al., 2021. Dual attention suppression attack: generate adversarial camouflage in physical world. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8561-8570. https://doi.org/10.1109/CVPR46437.2021.00846

Wei H, Tang H, Jia XM, et al., 2022. Physical adversarial attack meets computer vision: a decade survey. https://arxiv.org/abs/2209.15179

Wei XX, Pu BZ, Lu JF, et al., 2022. Visually adversarial attacks and defenses in the physical world: a survey. https://arxiv.org/abs/2211.01671

Wei XX, Huang Y, Sun YT, et al., 2023. Unified adversarial patch for cross-modal attacks in the physical world. Proc IEEE/CVF Int Conf on Computer Vision, p.4422-4431. https://doi.org/10.1109/ICCV51070.2023.00410

Wu T, Tong L, Vorobeychik Y, 2020a. Defending against physically realizable attacks on image classification. 8th Int Conf on Learning Representations.

Wu T, Ning XF, Li WS, et al., 2020b. Physical adversarial attack on vehicle detector in the CARLA simulator. https://arxiv.org/abs/2007.16118

Wu ZX, Lim SN, Davis LS, et al., 2020. Making an invisibility cloak: real world adversarial attacks on object detectors. 16th European Conf on Computer Vision, p.1-17. https://doi.org/10.1007/978-3-030-58548-8_1

Xu CJ, Ding WH, Lyu WJ, et al., 2022. SafeBench: a benchmarking platform for safety evaluation of autonomous vehicles. Proc 36th Int Conf on Neural Information Processing Systems, Article 1861.

Xu H, Ju A, Wagner D, 2021. Model-agnostic defense for lane detection against adversarial attack. Automotive and Autonomous Vehicle Security Workshop, Article 25.

Xu KD, Zhang GY, Liu SJ, et al., 2020. Adversarial T-shirt! Evading person detectors in a physical world. 16th European Conf on Computer Vision, p.665-681. https://doi.org/10.1007/978-3-030-58558-7_39

Xue MF, Yuan CX, He C, et al., 2021. NaturalAE: natural and robust physical adversarial examples for object detectors. *J Inform Secur Appl*, 57:102694. https://doi.org/10.1016/j.jisa.2020.102694

Yamanaka K, Matsumoto R, Takahashi K, et al., 2020. Adversarial patch attacks on monocular depth estimation networks. *IEEE Access*, 8:179094-179104. https://doi.org/10.1109/ACCESS.2020.3027372

Yang JH, Boloor A, Chakrabarti A, et al., 2020. Finding physical adversarial examples for autonomous driving with fast and differentiable image compositing. https://arxiv.org/abs/2010.08844

Yang KC, Tsai T, Yu HG, et al., 2021. Robust roadside physical adversarial attack against deep learning in Lidar perception modules. Proc ACM Asia Conf on Computer and Communications Security, p.349-362. https://doi.org/10.1145/3433210.3453106

Yang XH, Liu WF, Zhang SL, et al., 2021. Targeted attention attack on deep learning models in road sign recognition. *IEEE Int Things J*, 8(6):4980-4990. https://doi.org/10.1109/JIOT.2020.3034899

Yatsura M, Sakmann K, Hua NG, et al., 2023. Certified defences against adversarial patch attacks on semantic segmentation. 11th Int Conf on Learning Representations.

Yoon HJ, Jafarnejadsani H, Voulgaris P, 2023. Learning when to use adaptive adversarial image perturbations against autonomous vehicles. *IEEE Robot Autom Lett*, 8(7):4179-4186. https://doi.org/10.1109/LRA.2023.3280813

Yu Y, Yang WH, Tan YP, et al., 2022. Towards robust rain removal against adversarial attacks: a comprehensive benchmark analysis and beyond. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6003-6012. https://doi.org/10.1109/CVPR52688.2022.00592

Zeng FG, Dong B, Zhang YA, et al., 2022. MOTR: end-to-end multiple-object tracking with Transformer. 17th European Conf on Computer Vision, p.659-675. https://doi.org/10.1007/978-3-031-19812-0_38

Zhang H, Dana K, Shi JP, et al., 2018. Context encoding for semantic segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.7151-7160. https://doi.org/10.1109/CVPR.2018.00747

Zhang JD, Lou Y, Wang JP, et al., 2021. Evaluating adversarial attacks on driving safety in vision-based autonomous vehicles. *IEEE Int Things J*, 9(5):3443-3456. https://doi.org/10.1109/JIOT.2021.3099164

Zhang QZ, Hu ST, Sun JC, et al., 2022. On adversarial robustness of trajectory prediction for autonomous vehicles. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15138-15147. https://doi.org/10.1109/CVPR52688.2022.01473

Zhang Y, Foroosh H, David P, et al., 2019. CAMOU: learning physical vehicle camouflages to adversarially attack detectors in the wild. 7th Int Conf on Learning Representations.

Zhong YQ, Liu XM, Zhai DM, et al., 2022. Shadows can be dangerous: stealthy and effective physical-world adversarial attack by natural phenomenon. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15324-15333. https://doi.org/10.1109/CVPR52688.2022.01491

Zhu XP, Li X, Li JM, et al., 2021. Fooling thermal infrared pedestrian detectors in real world using small bulbs. Proc 35th AAAI Conf on Artificial Intelligence, p.3616-3624. https://doi.org/10.1609/aaai.v35i4.16477

Zhu XP, Hu ZH, Huang SY, et al., 2022. Infrared invisible clothing: hiding from infrared detectors at multiple angles in real world. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.13307-13316. https://doi.org/10.1109/CVPR52688.2022.01296

Zhu ZJ, Zhang YC, Chen H, et al., 2023. Understanding the robustness of 3D object detection with Bird'-Eye-View representations in autonomous driving. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.21600-21610. https://doi.org/10.1109/CVPR52729.2023.02069

Zolfi A, Kravchik M, Elovici Y, et al., 2021. The translucent patch: a physical and universal attack on object detectors. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15227-15236. https://doi.org/10.1109/CVPR46437.2021.01498

List of supplementary materials