Adversarial Attacks on Autonomous Driving Systems in the Physical World: a Survey

Lijun Chi, Mounira Msahli, Qingjie Zhang, Han Qiu *Member, IEEE*, Tianwei Zhang, Gerard Memmi *Senior Member, IEEE*, and Meikang Qiu *Senior Member, IEEE*

***Abstract*—Autonomous Driving Systems (ADS) represent a revolutionary advancement in transportation and offer unprecedented safety and convenience. Real-world physical attacks are emphasized because Autonomous Driving Systems (ADS) depend heavily on sensors and perception modules to detect and interpret their surroundings, making security a critical concern. Defenders usually have the upper hand in the digital sphere while they are challenged in the physical world because attackers have greater flexibility for covert operations. A comprehensive analysis is essential for understanding attack trends, evolution, and defense directions. This paper provides a survey of state-of-the-art physical attacks that threaten ADS perception. A novel multi-label classification method is introduced to categorize these attacks along four main dimensions. Visualization and analysis of the classification enhance the understanding of these multidimensional threats. Five research directions for future exploration are also proposed.**

***Index Terms*—Adversarial attacks, adversarial examples, black-box attacks, autonomous driving systems, environment perception**

I. Introduction

RECENTLY, Autonomous Vehicles (AVs) have been experiencing rapid development [1]. Benefiting from the advances of deep learning (DL), practical autonomous driving systems (ADS) are accelerating the commercialization of AVs. A typical ADS has a front-end perception module followed by an internal decision system [2]. For instance, Tesla's Autopilot uses cameras as its core sensor and relies on techniques such as occupancy networks for environment perception. Another approach, Waymo [3], not only uses cameras to capture visual contents but also uses LiDAR to present three-dimensional structural information of the target objects. Subsequently, the internal decision systems of modern ADSs utilize deep neural networks (DNNs) to process the information collected by these front-end sensors for decision-making.

Despite its promising performance, DNNs are well-known for their vulnerability to adversarial attacks [4]. In particular, attackers can carefully craft tiny (e.g. human-imperceptible) but malicious perturbations adding on the input sample (i.e. adversarial examples) to manipulate the DNN models' predictions. Today, there are various adversarial attack methods to generate adversarial examples (AE) in the digital domain [5]. The main focus of these AE-related studies is to generate tiny perturbations to keep the AE similar to the clean sample. For instance, AEs in computer vision (CV) tasks usually adopt a L_p bound to maintain the human-imperceptibility. Beyond this, the victim models also include natural language processing (NLP) [6], 3D point cloud [7], and speech recognition [8].

After digital adversarial attacks expressed a strong nuisance to DNNs [4], researchers began to explore whether such attacks would pose a risk to DNNs applied in physical scenarios. How to transfer digital perturbations to the real world was initially considered in CV tasks. Perturbations were most commonly printed as patches [9] or 3D objects [10]. Subsequently, this transfer was generalized to tasks such as 3D point clouds [11] or semantic segmentation [12]. With the high attack success rates gradually satisfied, researchers then tried to make attacks stealthier. For example, they made adversarial patches and objects seem like stains [13] and traffic cones [14]. Then, more flexible forms like lasers [15], and projections [16] were recently employed.

However, it is worth noting that physical adversarial attacks on ADS perception are significantly different from traditional physical adversarial attacks in the following three aspects. (1) **Internal sensors.** ADS contains sensors and models which attackers can target at the data collection and processing to mislead the decision bypassing the model. (2) **Heterogeneous data and models.** An ADS system contains both CV data and point cloud data which are heterogeneous, yet their models may perform the same tasks. Consequently, AEs for one model cannot mislead them both. (3) **Black-box DNN models.** Most current state-of-the-art (SOTA) adversarial attacks rely on a white-box scenario to calculate the gradients while DNN models in most commercial ADS are black-box.

In this paper, we aim to fully investigate the vulnerability of current ADSs based on an adversarial perspective. We focus specifically on published ADS attacks that have succeeded in the physical world. To comprehensively understand the current attack methods, we include those that can manipulate the ADS' decision targeting both DNN models and front-end components of ADS-like sensors. We also analyze several recently published papers at top-tier conferences in a quantitative way to illustrate the attack developing trends. In short, this survey makes the following contributions:

Lijun Chi, Mounira Msahli, and Gerard Memmi are with LTCI, Telecom Paris, Institut Polytechnique de Paris, Palaiseau, 91120, France. (email:{lijun.chi, mounira.msahli, gerard.memmi}@telecom-paris.fr).

Han Qiu and Qingjie Zhang are with Institute for Network Sciences and Cyberspace, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China, 100084. (email: qiuhan@tsinghua.edu.cn, zhangqingjie@sjtu.edu.cn).

Tianwei Zhang is with the School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore. (email: tianwei.zhang@ntu.edu.sg).

Meikang Qiu is with the School of Computer and Cyber Science, Augusta University, GA 30912, USA. (e-mail: qiumeikang@ieee.org).

- We investigate successful attacks on ADS systems in the physical world.
- We propose a quantitative analysis with visualization results to illustrate the developing attack trends.
- We review 50+ carefully selected physical adversarial attacks to analyze their methodology and results.

This paper is organized as follows: Section II presents the background of the ADS, which includes the workflow of the ADS and the detailed composition of the ADS system. Section III presents our method for analyzing, classifying, and visualizing selected papers. Section IV specifies the reviewed methods for physical attacks on the environment awareness in ADS systems. Section V discusses future research directions for ADS security. Finally, we conclude in Section VI.

II. Background

A. ADS working pipeline

Fig. 1 illustrates the workflow of a typical ADS. Information on the surroundings is first collected by sensors or cloud services. The sensors can be classified into motion-aware sensors, namely GNSS and inertial measurement unit (IMU), and environment-aware sensors such as LiDAR, cameras, etc. The collected information is then fed into the internal decision system (e.g. onboard DNNs) for processing. For instance, the position information can be combined with the high-definition map from the cloud servers to determine the vehicle's location. The environmental information is processed by onboard DNNs such as scene segmentation, object recognition, etc. to understand the vehicle's surroundings. Then, based on the above vehicle's positioning and surroundings, the planning stage determines actions to precisely control the chassis actuators. This paper focuses on attacks targeting the sensors as well as internal DNN models.

B. ADS perception systems

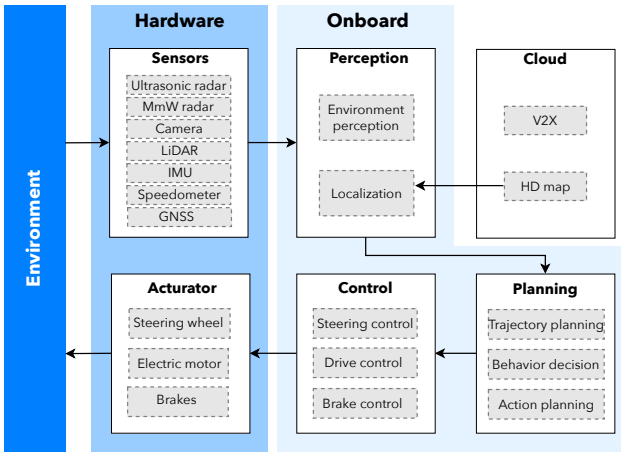


Fig. 1. ADS working pipeline.

1) Sensors of ADS: Most of the cameras contain vision-based sensors which are capable of taking colorful photos with rich textures. Obtaining high-quality photos is no longer a challenge for users because of the rapid development of camera hardware and image-processing technology. However, relying solely on 2D visual information from images is insufficient for self-driving cars, as the spatial position of obstacles is crucial beyond their mere presence. Additionally, poor lighting conditions degrade image performance. Consequently, special cameras have been developed to overcome these drawbacks. Space cameras such as monocular cameras can detect distant obstacles but in a narrow view. Binocular cameras calculate depth by parallax. And trinocular cameras enhance their view by combining three monocular cameras. On the other hand, no matter what the weather is like, thermal infrared cameras can perfectly capture heat radiation from humans and vehicles. Fig. 2 indicates that cameras are able to provide front, rear, side, and surround views.

A working LiDAR quickly rotates to scan its surroundings in a full circle with laser. When echoes bounce back, a 3D spatial map can be produced where each object around the vehicle possesses its own point cloud. ADSs can determine the current physical state of an object from the point cloud including its distance, trajectory, size, speed, and even shape. The precise information and the capability to operate in bright light environments are increasingly attracting ADSs to incorporate LiDAR into their frameworks. LiDAR's working area is represented by the orange range in Fig. 2. Indeed, the high price tag, commensurate with its superior capabilities, currently hinders widespread adoption. Moreover, rain, snow, and fog contribute a great deal of impurities to the medium, or air, through which signals are transmitted. The presence of these impurities may impair electromagnetic wave transmissions and lead to poor performance.

Ultrasonic radar relies on emitting high-frequency sound waves and analyzing the resulting echo to accurately measure short distances. Nevertheless, it is susceptible to temperature variations and exhibits reduced measurement precision when the vehicle is in high-speed motion due to the relatively slow propagation of sound. In autonomous driving, the sensor finds application in tasks like ascertaining unoccupied parking spaces and furnishing parking assistance, particularly within the ranges indicated by the red circle in Fig. 2.

Millimeter wave (mmW) radar obtains an object's distance by measuring the time it takes for the wave to bounce back and remains highly penetrative even in challenging conditions like rain and snow. Typically, the radar can offer accurate distance measurements up to 200 meters. Commonly used vehicle mmW radar is classified into two frequency bands: 24GHz and 77GHz. 24GHz mmW radar is used for close-range detection, such as blind spot monitoring and lane change assistance. 77GHz mmW radar is used for long-range detection, e.g. the distance and speed detection of the front car which is the basis for automatic emergency braking and adaptive cruise control. They work at the range defined by the green curves in Fig. 2.

The strengths and weaknesses of the sensors determine what perception tasks they are suitable for. Cameras excel at capturing color-based information, while LiDAR probes the surrounding spatial environment. Thermal infrared cameras enhance night vision, mmW radar specializes in long-distance detection, and ultrasonic radar facilitates close-range obstacle avoidance. Real ADSs often incorporate redundant sensors to ensure the robustness of autonomous driving across diverse

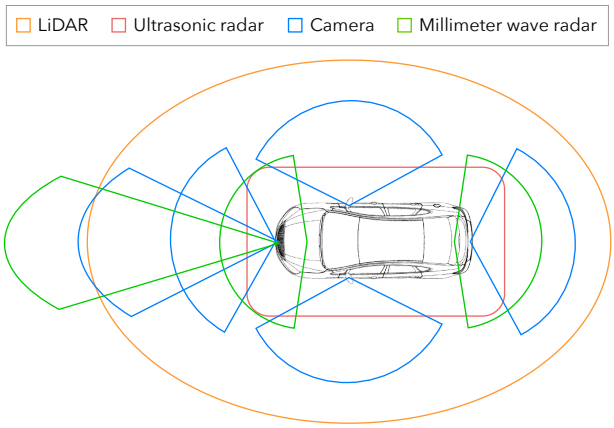


Fig. 2. Sensor working area.

environments. Moreover, sensor data that aligns and supports each other enhances credibility.

2) Internal DNNs: Much of the explosion in self-driving technology has come from the breakthroughs made by DL in the field of computer vision. Inspired by the human visual nervous system, Convolutional neural networks (CNNs) enable the reduction of volumes of data while preserving crucial features. The potential of this technique is being continuously explored with the support of hardware developments such as GPUs. So far, CNNs have made significant contributions to enhancing the accuracy of object detection and classification in images. The majority of advanced object detection methods rely on CNNs, and there are two main frameworks in this domain. First, single-stage detection, which enables both detection and classification within a single network, and second, region proposal networks (RPN) which are mainly composed of the generation of an overall region of interest (ROI) and the classification process.

AV necessitates reliable detection for accurate recognition of traffic signs, traffic lights, and road markings. The single-stage detection algorithm (e.g. YOLO series [17]–[20] and SSD [21]) offers fast inference and low storage costs, thereby making it well-suited for real-time autonomous driving applications. Conversely, region proposal detection frameworks such as R-CNN [22], Mask-RCNN [23], and Faster-RCNN [24] are more competitive in accuracy. However, they require substantial computational resources, which means that greater challenges are proposed in terms of training and fine-tuning. Maintaining a balance between performance and computational cost holds importance in autonomous driving. Since it is desirable to prepare enough reaction time for the planning and control module. In the current context, SSD is the preferred detection algorithm for ADS, but as hardware computing power continues to advance, RPNs are gradually being adopted in intelligent vehicles.

Dynamic driving activities, for example, obstacle avoidance cannot be accomplished without accurate physical features of surrounding objects. Nevertheless, acquiring target spatial information via vision methods refers to the problem of image matching from 2D to 3D, which brings a substantial computational overhead. Hence, many manufacturers choose to directly integrate LiDAR into their ADS. The point cloud they generate will then be employed for environment perception tasks. 3D CNNs constructed by referencing from vision methods have been applied to point cloud detection, such as the SOTA algorithms like PointPillars [25] and PointRCNN [26]. Despite their satisfactory detection capabilities, the massive computational volume of point cloud data throws strain on processing components. Methods based on depth maps or bird's eye views, such as PIXOR [27], have been put forward in this context for projecting point clouds onto a 2D plane.

Object tracking is another fundamental vision-based task equipped in self-driving vehicles. Tracking and monitoring the movement of objects of interest over time are the primary functions of this task. Various outstanding tracking algorithms have emerged from the vision-object-tracking (VOT) challenge. Some of the best models, such as SiamRPN [28] and DaSiamRPN [29], are based on Siamese networks. The crucial component is that they use RPN to identify possible regions where the object might be, and then introduce CNN to pinpoint the exact location of the target. Subsequently, trajectory prediction, which is usually performed at the back end of the perception module, utilizes the previously obtained tracking information for prediction. Notable algorithms in this area include GRIP++ [30] and Trajectron++ [31].

The previously mentioned detection networks focus on framing the target with a rectangle for classification, which overlooks relevant scene information. Therefore, semantic segmentation models [32] are developed to classify each input element (e.g., 2D pixels and 3D points). They generally use a combination of convolution for feature extraction and deconvolution for pixel-level labels [33]. Such networks perform well in drivable area segmentation and traffic line detection. Commonly used segmentation networks include PointNet [34], PointNet++ [35], SqueezeSeg [36] and Cylinder3D [37]. As such models require a great deal of computation to make predictions, they are slow, and, therefore, they are difficult to generalize to commercial ADSs.

C. Physical adversarial attack on ADS Perception

The traditional definition of AE involves perturbing pixels in the image to mislead the DNN (Section III-B2). In contrast, our interpretation of physical adversarial attacks has two main differences: (1) the attacker manipulates the physical scene without being restricted by pixels, and (2) the attack outcome relies on the victim ADS-perception system's response (sensors and DNN feedback). Fig. 3 illustrates the flow of physical adversarial attacks on ADS systems. The attacker performs unobtrusive and dangerous operations in the environment to disable sensors or deceive DNNs, which leads the ADS perception system to provide wrong perception results, such as incorrect traffic signs, traffic light colors, spoofed obstacles, etc. A driving plan is then developed based on inaccurate information during the planning phase. When the victim maneuvers according to the plan in the control phase, a traffic accident can occur.

1) Threat models: Three aspects are analyzed subsequently. **Attacker's goal.** An attacker intends to launch a stealthy attack against an autonomous driving vehicle in the physical world

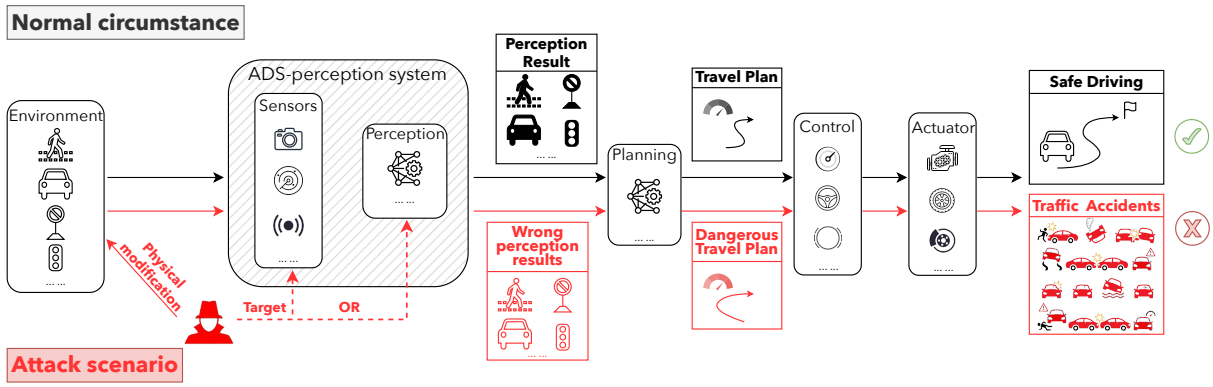


Fig. 3. Flow of physical attack.

in order to cause a potential traffic accident. Pursuing a high attack success rate involves two main challenges: remaining resilient to external factors (e.g., weather, light, angle) and avoiding disruption of human cognition.

Attacker's knowledge. The DNNs within the ADS of the victim vehicle are confidential, so typically the attacker does not know the detailed structure and parameters of internal networks. It is assumed that the attacker could take advantage of an open-source model that provides equivalent functionality. Aside from that, we assume that both the name and physical characteristics of the victim vehicle can be observed as semi-public information. Therefore, an attacker can obtain sensor parameters based on that information.

Attacker's capability. Besides assuming that attackers can perform attack scenarios in a simulator, we also consider whether attackers have sufficient financial resources to purchase identical sensor models or even to purchase vehicles. According to options previously discussed, the attacker may design and evaluate their attacks respectively on a simulator, a scaled-down model of the road, or a real road. We further assume that the attacker has the ability to modify the real environment via several common methods, including placing objects, pasting paper, or arranging lighting.

2) Setup: Attackers consider the following three factors.

DNN models. Some physical attacks systematically control ADS behavior to produce attackers' desired outcomes in the targeted task. Thus these attacks can be evaluated using the victim models corresponding to the targeted DNNs. Aside from the networks discussed in Section II-B2, state-of-the-art classifiers such as VGG-ens [38], ResNet-ens [39], DenseNet [40], and Inception v3 [41] can serve as victim models, as they are the classification phase within the RPN system.

Datasets. Datasets for training or attacking environment-aware models can currently be divided into two categories: 2D image datasets from cameras and 3D point cloud datasets from LiDAR and millimeter wave radar. The first category consists of photographs of traffic signs in various natural environments such as GTSRB [42], LISA [43], and TT100k [44]. For the extended field of view, datasets such as BDD100K [45] provide live maps, videos, and images for driving on lanes and routes. ImageNet [46] and COCO [47] are collections of images for image classification. In the second category, KITTI [48] is a widely used point cloud dataset with up to six hours of real-time traffic data. Furthermore, nuScenes [49] offers point clouds, surrounding images, and driving information for exploring sensor fusion.

ADS. Open-source autonomous driving platforms such as Baidu Apollo and Openpilot are commonly chosen for the experiment by being loaded and run on mobile devices. CARLA [50], built on Unreal Engine 4, offers a wide range of high-quality maps and vehicle models. The attacker adjusts various elements within the simulator, including vehicles, textures, and camera states, to accurately simulate an attack scenario. Compared to purchasing highly autonomous vehicles such as Teslas, simulators and other open-source ADSs offer cost advantages and a greater number of trials, making them more attractive to attackers.

3) The challenges: Executing digital attacks successfully in the physical environment while ensuring their effectiveness is an enormous challenge. A number of external factors can influence the success rate of an executed attack, including color loss in sticker printing, incorrect perturbation placement, changes in illumination intensity, changes in target angle during vehicle movement, changes in background color, etc. [51]. In an effort to enhance the robustness of attacks in both the digital and physical domains, attackers have used methods such as Expectation over transformation (EOT), first proposed by Athalye et al. [52]. EOT contributes to resilience against camera-induced changes, ensuring that physical perturbations remain adversarial even when subjected to transformations such as changes in distance, angle, exposure, etc. As well, Evtimov et al. [53] proposed that traffic signs obtained from multiple real-world environments could be utilized as inputs to the model during training. NatureAE [54] maintains a high attack success rate (ASR) while minimizing pixel perturbations to maintain the stealthiness of the physical sticker during training. In addition, considerable research efforts have been devoted to the direct design and execution of physical-world attacks recently to eliminate inconvenient upscaling [15].

D. Comparison with existing survey

Several surveys have looked into the complete range of autonomous driving security issues. Qayyum et al. [55] investigated security threats arising from machine learning in

TABLE I
COMPARISON WITH EXISTING SURVEYS.

Surveys	Physical black-box attacks on ADS perception	Year	Number
Qayyum et al. [55]	✓	2020	5
Deng et al. [56]	✓	2021	16
Woitshcek et al. [57]	✓	2021	5
Mahima et al. [58]	✓	2021	6
Gao et al. [59]	✓	2021	9
Wei et al. [60]	✓	2022	7
Ours	✓	2024	**51**

connected and autonomous vehicles (CAVs), particularly adversarial attacks. Deng et al. [56] discussed both white-box and black-box attacks on autonomous driving that compromise ADS from the inside and outside. Woitshcek et al. [57] investigated the feasibility of physical adversarial attacks in traffic sign recognition. Mahima et al. [58] focused on adversarial attacks targeting object detection, classification, and semantic segmentation in self-driving. Gao et al. [59] summarized the security challenges faced by the various systems making up ADS. Wei et al. [60] reviewed adversarial attacks on computer vision tasks in the physical world. It can be seen from Table I that physical attacks on black-box ADS perception are only a small part of the security threat research. With the constant updating of attack techniques in recent years, this topic has evolved into a significant and unavoidable field. There is an urgent need for comprehensive research on physical adversarial attacks on ADS perception.

The research described above demonstrated that attacks can be classified from various perspectives. However, they dwell on classical AEs in categorization and fall short of addressing the diversity of physical attacks nowadays. In light of this, an innovative multi-tag classification strategy is proposed in our survey. For instance, Deng et al. [56] define attacks on sensors as physical attacks and those on DNNs as adversarial attacks. In our threat model, a physical attack refers to an attack occurring in the real-world environment. Therefore, sensors and DNNs serve as the sub-tags of the main factors of attack generation. Gao et al. [59] described attacks based on victim sensors. However, the same sensors with different perception models can perform various tasks. Thus, we define a more detailed sub-tag called attack task. Wei et al. [60] categorize adversarial attacks based on tasks and forms, but their sub-tags are not comprehensive (5 attack forms and 6 attack tasks). We supplement both tags with more categories (8 forms and 11 tasks). The classification approach enables an accurate description of the attack method in the physical world while maximizing the inclusion of information by rationally defining relevant tags. Fig. 4 depicts the pipeline of research encompassing literature collection, label design, numerical analysis to assess the validity of label combinations, comprehensive analysis of attack methods, and future predictions.

III. ATTACK CLASSIFICATION

A. Distribution of papers

The main challenge of literature research is to know the distribution of abundant existing papers and synthesize a classification standard. Nevertheless, the methodology of literature research remains in a literal way, ignoring data analysis skills. In this section, we propose a novel methodology to visualize the distribution of papers and reveal their connections, which could furthermore help us determine the importance of classification standards.

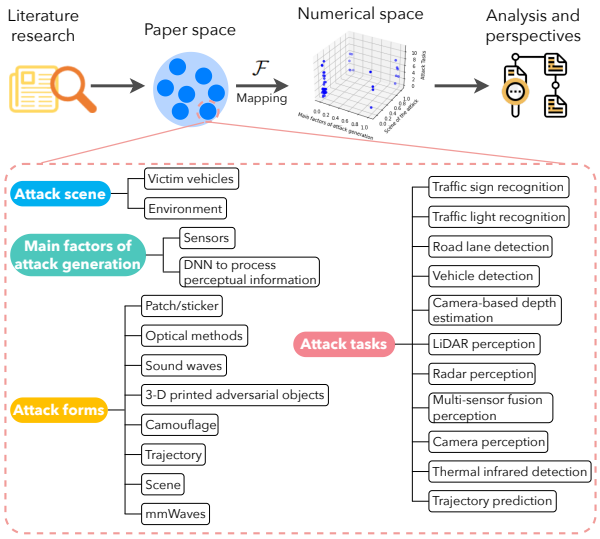


Fig. 4. Pipeline of our work.

After summarizing all papers and literally analyzing their similarities and differences, we generally induce four classification standards (introduced in Section III-B): attack scene, main factors of attack generation, attack tasks, and attack forms. Each paper can be classified into different clusters based on various classification standards, allowing it to be represented effectively by four labels. In order to visualize the distribution of papers, we propose to construct a data set of multi-label classification [61]. Let $\mathcal{X} = S \times M \times T \times F$ be the space of papers, where S (resp. M, T, and F) is the space of Attack scene (resp. Main factors of attack generation, Attack tasks, and Attack forms). Let $\mathcal{Y} = [0, 1, ..., \mathbf{card}(S) - 1] \times [0, 1, ..., \mathbf{card}(M) - 1] \times [0, 1, ..., \mathbf{card}(T)-1] \times [0, 1, ..., \mathbf{card}(F)-1]$ be the numerical space to visualize data distribution, where $\mathbf{card}$ denotes the cardinality function. We map each paper $x \in \mathcal{X}$ to a position $y \in \mathcal{Y}$ through a mapping function:

$$\begin{aligned} &\mathcal{F} : \mathcal{X} \to \mathcal{Y}, \\ &(s, m, t, f) \to (\mathbb{I}(s), \mathbb{I}(m), \mathbb{I}(t), \mathbb{I}(f)) + Gauss, \end{aligned} \tag{1}$$

where $\mathbb{I}$ denotes the index function and $Gauss \sim \mathcal{N}(0, 0.02)$ denotes the Gaussian noise which differentiates papers on the same position.

Since we cannot visualize 4-D data in a 3-D space, we choose three labels out of four and iterate the visualization four times. Fig. 5 shows our visualization results, from which we summarize four findings:

- The upper two sub-figures show a chaotic distribution. This is attributed to the wide diversity of attack tasks and forms covered in the papers, encompassing 11 attack tasks such as

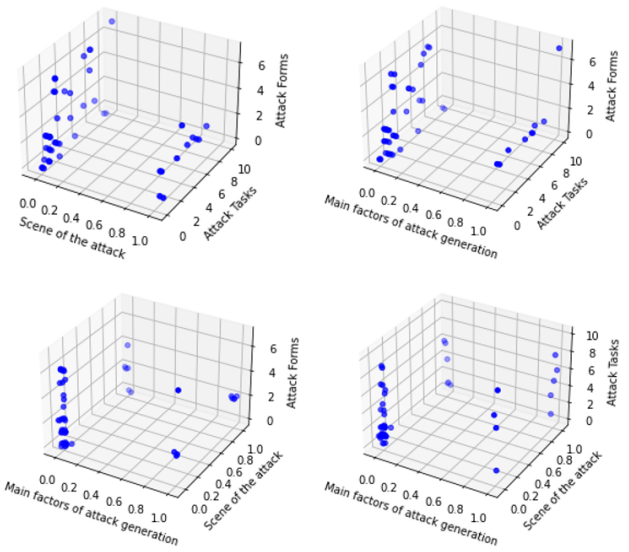


Fig. 5. Visualization of the distribution of papers.

traffic sign recognition and vehicle detection, and 8 attack forms including patch-based and sound wave attacks.

- The lower sub-figures reveal a clearer distribution with four distinct vertical clusters. It highlights the significance of classifying attacks based on an attack scene and main factors of attack generation. These two factors are of primary concern for attackers in real-world scenarios.
- Among the bottom two sub-figures, the right one appears neater, suggesting that attack tasks are more influential than attack forms. When the vertical axis shifts from attack forms to attack tasks, the four clusters exhibit a similar structure, which results in a more balanced distribution of papers across attack tasks.
- To determine the order of importance between the attack scene and main factors of attack generation, we refer to the upper sub-figures where we observe two clusters. The upper left cluster, which is less imbalanced, suggests that the attack scene holds greater importance. The clusters' balance, indicating the equal distribution of papers, highlights the significance of the classification standard.

When classifying a paper on black-box physical attacks on ADS, priority should be given to evaluating its attack scene, followed by its primary factors of attack generation, then its attack tasks, and finally its attack forms.

B. Classification via Multi-tags

For the purpose of delineating various physical attack methods comprehensively, a classification consisting of four primary tags is proposed. The framework gives us a four-dimensional view of the attack. The following part of this section explains and defines each of the four primary categories proposed in our methodology.

1) Attack scene: Our first point of interest lies in determining which part of the scene interacts with an attacker in the real world. As a general rule, the methods and tools employed to commit a physical attack depend on the attack scene. Two sub-tags are therefore defined: the victim vehicle and their environment.

Victim vehicles. Interaction with a compromised vehicle essentially refers to dealing with a variety of sensors with their own unique function. One noteworthy feature of these attacks is that they seek to contaminate the input data of the sensors by manipulating the signals. It is achieved while the real environment remains unchanged. Taking a visible light camera as an example, a malicious attacker might change the trajectory and color of light coming into a visible camera. Beyond that, fraud lasers or mmWave signals may be used to disrupt data collection by LiDARs and mmW radars.

Environment. The physical environment serves as the source of information that enables an attacker to manipulate elements surrounding the victim vehicle and update the information obtained at its source. The fundamental principle behind this tag is proactive engagement with the surroundings to simplify operational procedures and enhance covertness strategies. In light of this, most of the entities selected for interaction are of strategic importance in the field of traffic regulations, especially traffic signs, traffic lights, and road lanes. As an added bonus, attackers can strategically position themselves as participants in traffic accidents by inducing the victim vehicles to collide with them or directing these vehicles onto incorrect routes. Several studies have explored staging attacks in common settings such as open airspace or ordinary roads to increase the flexibility of attack activation.

2) Main factors of attack generation: As stated in Section II-B, the environmental awareness of the ADS consists of two components, the sensor and the internal network. So the second tag is defined to specify the target components during physical attack designing and highlight the main factors driving attack generation.

Sensors. Although sensor technology continues to be refined to meet human requirements, some inherent vulnerability still remains. Sensors like LiDAR and millimeter wave radar actively engage with their surrounding by emitting signals and analyzing the returned data. However, the receiver has no way of verifying the authenticity of the data. Attackers can therefore fool the sensors by sending out deceptive signals through suitable signal-emitting devices. Conversely, cameras passively acquire environmental light and possess the ability to detect frequencies extending beyond human vision (380nm to 700nm). The lens flare phenomenon occurs when strong light directly enters the camera lens. The incoming light (red line) reflects off elements, like apertures, lenses, etc., and introduces artifacts that reduce image quality (as shown in Fig. 6).

DNN to process perceptual information. When changes based on gradient orientation occur, it becomes challenging to maintain high accuracy even if the changes are minor. This kind of attack is known as an adversarial attack which has been a significant concern ever since it first emerged. Recent research has revealed that it is difficult to resist adversarial attacks no matter how complex the DNN structure. An ADS perception DNN can be expressed as $F(x;\theta) = y$, where θ is a model parameter and y is the output of the clean sample x. The corresponding adversarial example (AE) is denoted as $\widetilde{x} = x + \delta$, where δ is the adversarial perturbation. The

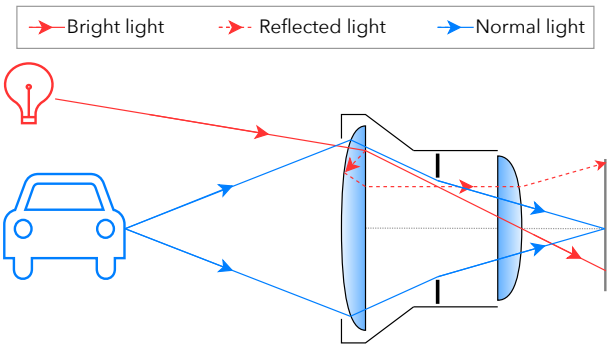


Fig. 6. Lens flare.

aim is $F(\widetilde{x};\theta) = t$ in targeted attacks and $F(\widetilde{x};\theta) \neq F(x)$ in untargeted attacks. Essentially, the hope is that the target model will produce an incorrect output. DNN-based attacks can usually be formulated as optimization problems of the following form [4].

$$\begin{aligned} \min_{\delta} \quad & L(F(\widetilde{x};\theta), t), \\ \text{s.t.} \quad & d(x,\widetilde{x}) < \varepsilon, \end{aligned} \tag{2}$$

where L denotes the loss function. δ is constrained designed and ε is the distance-related hyper-parameter.

3) Attack tasks: The third principal tag is formulated based on the compromised functions within the ADS. A significant amount of authentic data is required by the decision-making phase in ADS. Complex perception functions that may even involve redundant tasks across various sensors are necessary to ensure data accuracy and reliability. These functions are systematically organized into the following 11 sub-tags.

Traffic sign recognition. The key information ADS seeks in this task is the location and type of traffic signs. ADS will obtain the most current information about traffic regulations after traffic sign recognition. With the advent of CNNs, this CV-based task has taken a giant leap forward [62].

Traffic light recognition. This mission aims to pinpoint the locations of traffic lights and distinguish their color statuses. Traffic lights need to be seen by the ADS camera and passed to the perceptual network [63]. Depending on the color displayed by the relevant traffic light, the ADS will adapt its following behavior accordingly.

Road lane detection. This detection task allows self-driving vehicles to stay centered in their lane or change lanes smoothly [64]. White or yellow lines are noticed everywhere on roads which divide the road into specific zones. Each of them remarks a safe space where vehicles and pedestrians can move without interfering with each other. For self-driving vehicles, the equipped ADS needs to detect these lines using either a camera or LiDAR.

Vehicle detection. Sensors such as cameras, LiDAR, and millimeter-wave radar are utilized to locate nearby vehicles [65]. Most roads are public spaces shared by all users, which means that vehicles will typically encounter other vehicles. Autonomous vehicles are expected to be aware of their neighbors so as to avoid potential collisions between them. Whether cruising or parking, the task guides the vehicle in planning an appropriate route.

Camera-based depth estimation. Photographs acquired by monocular [66] or stereo cameras [67] are pertained to estimate depth information. Some manufacturers select the most cost-effective cameras to gather spatial depth data and measure the distance to the vehicle or obstacle in front of them. This task can be employed to ensure the vehicle maintains a secure distance from the vehicle in front.

LiDAR perception. The purpose of this task is to provide spatial features, such as location, distance, shape, etc., of surrounding objects. ADSs can build a 3D point cloud to measure these objects as precisely as possible with LiDAR technology [68]. As a result, self-driving vehicles can safely navigate around obstacles by sensing everything around.

Radar perception. Critical information such as the precise distance, shape, and size of objects within a short distance is basically offered by this task [69]. MmW radar effectively creates a sparse point cloud essential for neighborhood perception and is commonly used in assisted parking.

Multi-Sensor Fusion (MSF) perception. This consists of prior detection tasks but requires sensor collaboration to enhance and validate the final outcomes [70], [71]. Camera and LiDAR fusion, for example, can be used to identify the type of obstacle according to its visual and physical features.

Camera perception. This covers traffic sign recognition, lane detection, vehicle detection, etc. Although these tasks are used in different ADS applications, they all suffer from the same camera weaknesses.

Thermal infrared detection. The infrared camera is capable of capturing heat-based images [72]. These images facilitate the detection of heat-emitting subjects such as people and animals during nighttime and adverse weather conditions.

Trajectory prediction. Real-time targeting of pedestrians or adjacent vehicles is performed using cameras or LiDAR sensors during driving. The tracking outcomes [73] are useful to anticipate the target's future trajectory [74]. This sub-tag contains both tracking and prediction.

4) Attack forms: The fourth tag corresponds to the available physical attack forms. Unlike traditional adversarial attacks in the digital world, physical attacks cannot inject pixel-level perturbations into reality. Most physical attacks take advantage of natural phenomena or attachments to achieve adversarial perturbations. The following sub-tags are defined based on the analysis of attack techniques.

Patch/sticker. Using a patch or sticker is a classical method of upscaling the digital world adversarial attacks to reality. It involves training the perturbed pixels in a regular small area and then creating a 2D patch or sticker in the real world for application to the designated location.

Optical methods. These allow attackers to meddle with captured images without touching anything. Increasing, decreasing, or changing the light in the scene are three basic attack approaches. Applying specific emitters, different types of electromagnetic waves can be added to the camera. Decreasing light is achieved by putting a barrier between the light source and the object to create shadows. Attackers can change the lighting depending on optics as well through lenses or mirrors.

Sound waves. An acoustic source producing sound waves triggers molecular vibrations. When the energy reaches a certain threshold, the vibration can be transmitted to other objects and cause mechanical oscillation. This form of attack uses sound waves that are strong enough to shake the sensor resulting in fuzzy inputs.

Adversarial objects. A 3D adversarial sample is acquired during training and subsequently produced using a 3D printer. Upon fabrication, attackers can position the object at a specific location in the physical environment.

Camouflage. Spray paint and graffiti strategies are ideal methods for camouflaging existing patterns as well as painting new facades. Camouflage should be designed considering the curvature of the object's surface and then the attacker can reproduce it in the target area.

Trajectory. The trajectory can be either the victim vehicle's moving path or the trajectory of a participating vehicle controlled by the attacker.

Scene. Establishment of a scene requires the cooperation of all subjects within the scene. It involves the strategic placement of participants and objects in predetermined locations.

MmWaves. Certain devices have the capability to accurately simulate mmW radar echoes and subsequently transmit fabricated signals to a target receiver.

The choice of attack form directly affects its characterization and implementation. To systematically evaluate the potential risks raised by various attack forms, we propose four key criteria: feasibility, visibility, the flexibility, and expense.

Feasibility measures the simplicity of attack implementation. Attack forms are considered highly feasible if they require minimal equipment and shorter installation time, such as patches, optical methods, etc. In contrast, altering a vehicle's trajectory without firm plans and setting up complex scenes with multiple vehicles and obstacles are classified as low feasibility due to their difficulty and extensive resources needed.

Visibility marks how noticeable a perturbation is. High visibility attacks, such as patches, adversarial objects, and camouflage, are easily detected by humans. In contrast, attacks that hide their source, such as optical methods or those that manipulate the original objects in the environment, such as trajectory and scene, are low visibility attacks.

Flexibility evaluates how easily the attack's activation and location can be controlled. Methods that can be activated at any time and move the attack's location have high flexibility, which is not the case for patches, camouflage, and scene. Low flexibility attacks, such as patches and scene manipulations, are more context-specific.

Expense is the cost of the equipment required for the attack. We looked up the equipment prices mentioned for each form and categorized expenses as follows: less than 100 dollars as low-cost, between 100 and 1000 dollars as medium-cost, and over 1000 dollars as high-cost. However, for certain trajectories, implementation plans are not yet available. The cost of optical methods varies widely. For instance, using a wave generator to simulate LiDAR is expensive, while tools such as laser pointers are more affordable.

Attackers aim to make their methods more dangerous and easily achievable by ordinary individuals without requiring extensive knowledge, equipment, or resources. Thus, the trend in attack evolution is toward simpler setup, greater stealth, and lower cost. This corresponds to our characterization of high feasibility, low visibility, high flexibility, and low expense. Table II illustrates these criteria across various attack forms. Traditional attack forms such as patches and adversarial objects tend to be high visibility and low expense, while new attack forms such as signal simulations (LiDAR, sound waves and mmWaves) and scene are low visibility and high expense. It shows that the variety of attack forms has increased in recent years, particularly in terms of greater feasibility, visibility, and flexibility. Whereas there is no significant advantage of these forms in terms of cost. The first three criteria are often prioritized when designing an attack, which can lead to increased costs. However, with recent improvements in these three metrics, the new focus is on reducing the expense of attacks. As can be seen from Table II, the only form meeting all these requirements is the optical method without LiDAR simulation. This is why such attack forms have flourished in recent years.

IV. Analysis of Attack Classification

For the purpose of this survey, 51 articles from prominent conferences such as USENIX Security, ACM CCS, IEEE Oakland, NDSS, and CVPR, covering the period from 2020 to 2024, focusing on published SOTA physical attack methods were gathered and organized in Table III all the while utilizing the multi-label approach. The analysis reveals the following key insights.

- Traffic sign recognition continues to be the task that is most commonly targeted in ADS.
- Optical means are a popular choice for attacking sensors with only one instance utilizing mmWaves to generate false signals.
- Attacks on thermal infrared detection, trajectory prediction, and vehicle detection have increasingly relied on DNNs and environmental interaction.
- Recent attacks on traffic light recognition are primarily sensor-based and interactive with victim vehicles.
- Most techniques designed to exploit sensor vulnerabilities typically require interaction with the victim vehicle, except for phantom attacks [67].

Since LiDAR and mmW radar are both electromagnetic-wave-based sensors, optical methods have been a mainstream form of attack against them. Cameras, due to their imaging principles, have also become the latest victims (see Section IV-B). Consequently, optical methods have gradually become a popular trend in attacks. Sensor-based attacks prefer direct interaction with sensors rather than indirect effects through the environment because direct contact theoretically reduces energy loss and the risk of being pre-detected by humans.

Attacks with the same generation factors prefer using similar techniques. Therefore, our analysis is segmented into two main subsections: DNN-based attacks and sensor-based attacks. We discuss common applicable techniques before addressing specific forms and tasks for each attack classification. And the attack scenes usually shape the layout of the attack

TABLE II
CHARACTERISTICS ANALYSIS OF VARIOUS ATTACK FORMS.

Attack forms / Features	Patch/ sticker	Optical methods	Sound waves	Adversarial objects	Camouflage	Trajectory	Scene	MmWaves
Feasibility	High	High	High	High	High	Low	Low	High
Visibility	High	Low	Low	High	High	Low	Low	Low
Flexibility	Low	High	High	High	Low	High	Low	High
Expense	Low	Low/High	High	Medium	Low/Medium	Unknown	High	High

forms, which is also mentioned later. Furthermore, we provide a brief overview of common defense mechanisms and some challenges they face.

A. DNN-based attack

How the loss function L is designed and solved is the key to a DNN-based attack (as mentioned in Section III-B2). The L typically aims to minimize the performance of the victim model while meeting the requirements of the scenario. Traditional optimal solutions rely on white-box model gradients such as projected gradient descent (PGD). PGD algorithm iterates over AEs according to the gradient of L to maximize the value of the loss function itself. These AEs likewise demonstrate their attack potential against black-box DNNs [75], known as transfer-based attacks. Current research improves the AE transferability by designing diverse loss functions such as attention attacks [76] and public attention attacks [77], in which the loss function aims to shift model attention or public attention to the wrong class.

Some query-based attacks simulate the victim model or approximate gradient values from input-output results so as to employ white-box optimization techniques [78]. Alternatively, query-based AEs can also be generated by observing the outputs alone. This goal can be achieved through the genetic algorithms [79], particle swarm optimization (PSO) [80], or grid search [81] during training. Genetic algorithms assign features to genes in the chromosome and optimize solutions by simulating biological evolution processes such as heredity, mutation, and selection. PSO mimics the behavior of a flock of birds searching for food where each bird continually moving to seek the location with the most food overall. In other words, each particle updates both its speed as well as position to maximize the loss function. Grid search is suitable for cases with a limited number of parameter combinations. The optimal parameter combination is found through exhaustive exploration. Physical attacks provide attackers greater flexibility in employing innovative attack forms, meaning that more appropriate loss functions need to be devised. And optimization problems are often solved using the classical algorithms described above. Thus, we analyze the objectives that each loss function aims to achieve and specify the optimization algorithm selected for this purpose.

Attack LiDAR perception by adversarial scene. Attackers have the ability to create scenes specifically designed to deceive point cloud-based segmentation [32]. A tree structure was utilized to represent scene data, where nodes represented objects and edges depicted the relationships between them. Their goal of the optimization is to search for a tree structure to minimize model performance while ensuring compliance with the rules, with the potential search space being provided by the VAE model [82]. The authors successfully steered victim vehicles into collisions by setting up the environment in the CARLA simulator according to the adversarial tree structure.

Fig. 7. Attacking LiDAR perception by 3D object (image credit: [83]).

Attack LiDAR perception by adversarial objects. The geometry of the adversarial object is determined by a collection of points and surfaces arranged in a mesh. The attacker continuously adjusted these elements through a genetic algorithm [79] and then employed a 3D printer to produce the adversarial object. The object was strategically placed within the field of vision of the targeted vehicle. From Fig. 7 we can visually observe that in Yang et al.'s study [83], they trained and produced a polyhedron that the LiDAR system perceives as a vehicle to be avoided.

Attack MSF perception by adversarial objects. The MSF algorithm integrates camera and LiDAR inputs to formulate predictions. An effective attack would therefore need to impact both the image and point cloud perception networks in a manner. To execute such an attack, attackers could generate a mesh 3D, print it, and strategically place it. Utilizing the gradient of the open-source MSF, PGD can be employed to train the adversarial mesh. Cao et al. [14] devised an adversarial traffic cone that effectively avoided MSF detection. Their experiment involved a miniature road model (depicted in Fig. 8 (a)) with a cell phone and a LiDAR capturing images and point clouds. The MSF presented in Fig. 8 (b), (c) was incapable of identifying the adversarial cone both visually and in the point cloud.

Attack camera-based depth estimation by patch. The attacker strategically placed an adversarial patch at the rear of the targeted vehicle and caused the following victim vehicle to perceive inaccurate spacing. Cheng et al. [84] devised

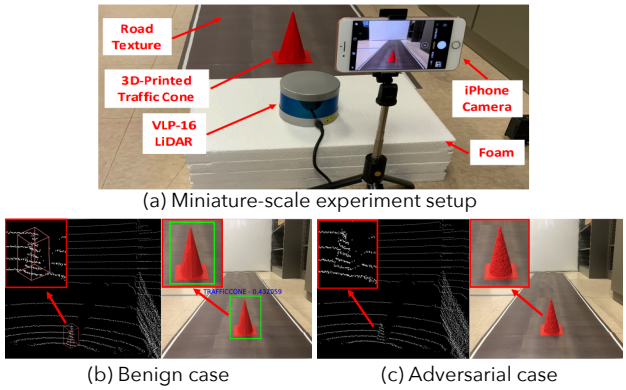


Fig. 8. Attack MSF perception by 3D object (image credit: [14]).

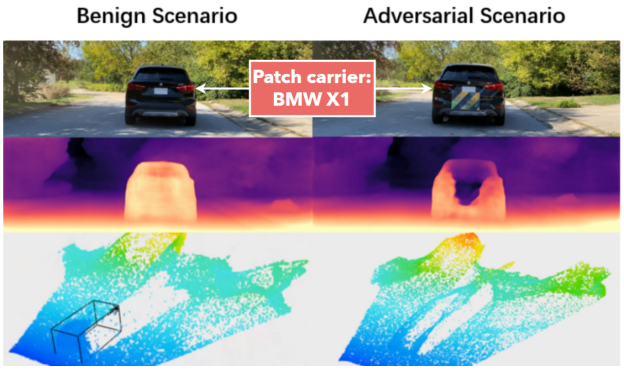


Fig. 9. Attack camera-based depth estimation by adversarial patch (image credit: [84]).

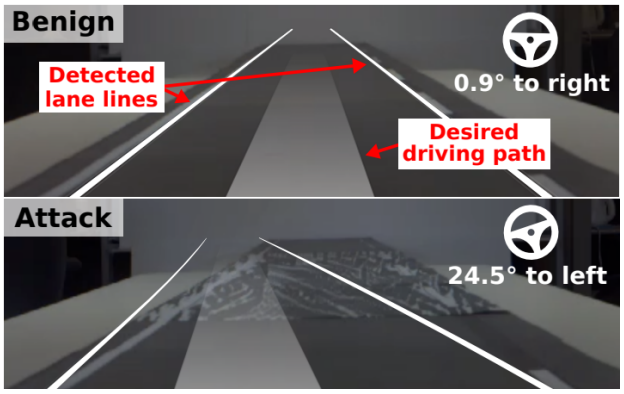


Fig. 10. Attack Road lane detection by patch (image credit: [13]).

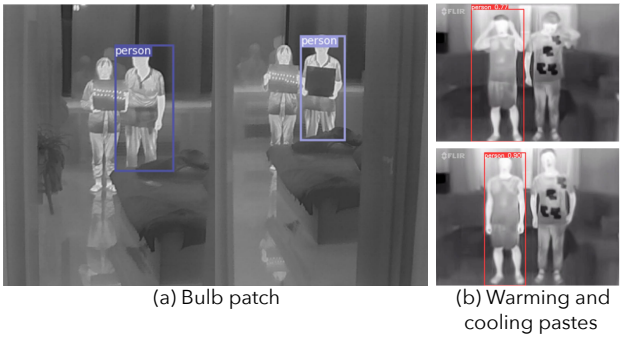


Fig. 11. Attack thermal infrared detection by heat adversarial patch (image credit: [88], [89]).

the patch to amplify the attack's impact by placing smaller patches in highly responsive regions. The optimization process employed the gradient of the white-box depth estimation model and LBFGS algorithm [85] to generate transferable adversarial patches. As illustrated in Fig. 9, the patch carrier drives in front of the victim while the victim ADS fails to detect this car.

Attack road lane detection by patch. Road lane detection attacks are often carried out by placing patches on the road. These patches cause the ADS to incorrectly identify the lane when the victim vehicle passes over them, which can lead to steering at the wrong time. Boloor et al. [86] adopted colored lines attached to the road surface for their attack. As collisions are dynamic, they considered the effectiveness of the attack over consecutive frames and formulated the optimization problem accordingly. The researchers employed a grid search method to navigate through different possibilities so as to find the solution. The line patches effectively induced missteering in a CARLA. Jing et al. [87] employed a heuristic algorithm to determine the query-based perturbation. They mapped their patch on a real-world road to execute an attack on a Tesla Model S. In contrast, Sato et al. [13] devised a transfer-based perturbation and used a more natural representation, road dirt, as shown in Fig. 10.

Attack thermal infrared detection by patch. Alterations in local temperature can introduce perturbations to thermal imaging so that accurate ADS pedestrian detection at night becomes impossible. Zhu et al. [88] developed a transfer-based attack that takes advantage of pixel changes induced by radiant heat from illuminated bulbs. They discovered the best positions for small bulbs on a magnet plate using PGD algorithms. The resulting arrangement, as depicted in Fig. 11(a), was practically applied to create thermal patches. Nevertheless, the visibility of handheld bulb plate in darkness encouraged the exploration of covert manipulation methods. Wei et al. [89] proposed the application of warming and cooling patches to generate anomalous color patches on human thermal imaging, as demonstrated in Fig. 11(b). Invisibility was achieved by affixing patches to the interiors of clothing. They optimized the attributes of these patches with the PSO algorithm. Both approaches result in pedestrians becoming undetectable when their heat maps are processed by the detection network.

Attack traffic sign recognition by patch. Using patches or stickers to manipulate traffic sign recognition is a well-known physical attack method. Wei et al. [90] improved the fixed-location sticker-based attacks by adding positional parameters of the stickers to the optimization objective. The modification aimed to place the stickers in highly sensitive regions for a greater impact. They also developed another attack that utilizes existing patterns as perturbations and optimized parameters such as position, size, and rotation angle [91]. Both attacks, depicted in Fig. 12(a), can lead to the failure of traffic sign recognition. Additionally, Giulivi et al. [92] employed scratches as perturbations in their query-based attack depicted in Fig. 12(b). Zhu et al. implemented

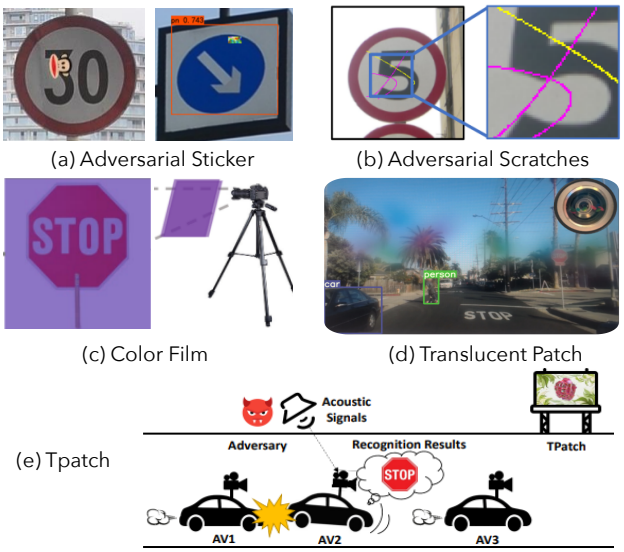


Fig. 12. Attack traffic sign recognition by patch (image credit: [91], [92], [90], [95], [94], [96]).

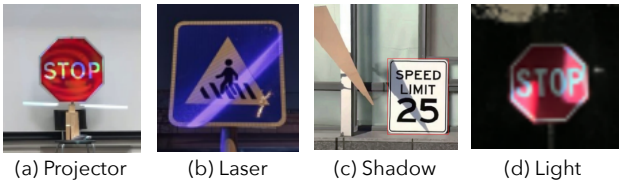


Fig. 13. Attack traffic sign recognition by optical methods (image credit: [99], [15], [102], [103].

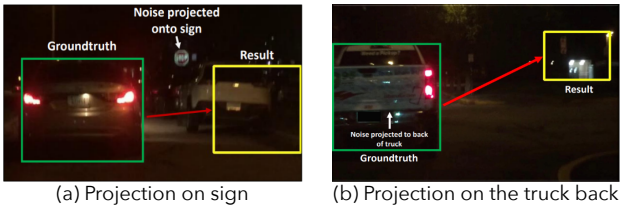


Fig. 14. Attack trajectory prediction by projector (image credit: [105]).

an adversarial sticker activated by a specific sound signal to improve the controllability of the sticker attack (Fig. 12(e)). They constructed a model to replicate the camera jitter image caused by sound. Employing gradient optimization, the attack intended to transform the post-blur perturbation to a stop sign while maintaining a conventional pattern before blurring. The sound interacted with the sensors but the attack fell under the tag of environment due to a manipulation introduced into the surroundings. New materials allow the patches to get novel capabilities. Tsuruoka et al. [93] utilized reflective patches to disrupt traffic sign recognition models at night by making them visible under the illumination of vehicle headlights. The reflections were represented as white squares and the attacker sought optimal grid locations to position the patches.

Transparent stickers stuck to the camera can also disrupt traffic sign recognition. Fig. 12(c) illustrates how Hu et al. [94] used color films to simulate haze effects. They trained various film parameters, such as color, transparency, and other pertinent physical attributes, through a genetic algorithm. Instead of the obvious monochromatic film, Zolfi et al. [95] developed a transferable translucent sticker via gradient optimization techniques. They situated an attacked camera in front of a playback driving video and analyzed how the physical attack affected the detection frame. In Fig. 12(c) the traffic sign is lost in detection.

Attack traffic sign recognition by camouflage. Almost all camouflage attacks need perturbations to be fully overlaid on traffic signs. Over the last three years, researchers in this field have been working on improving attack transferability. Xue et al. [54] and Jia et al. [97] introduced the alteration of factors during training such as background and target sign orientation to raise perturbation robustness. Yang et al. [98] proposed an optimization approach based on a soft attention graph to enhance attack transferability.

Attack traffic sign recognition by optical methods. In comparison with static stickers, dynamic optical techniques have gained significant popularity in recent years. A range of economical optical devices are employed to manipulate traffic sign recognition. Gnanasambandam et al. [16] and Lovisotto et al. [99] adopted a projector to project adversarial perturbations onto traffic signs to disrupt the ADS detection phase, as illustrated in Fig. 13 (a). Xie et al. [100] proposed a frame-by-frame projection attack, considering the ongoing dynamic process of a victim vehicle approaching.

Monochromatic light serves as another form of physical perturbation. Duan et al. [15] introduced an attack utilizing a monochromatic laser beam depicted in Fig. 13 (b). Different forms of light, like spots, beams, and light zones, are employed by Hu et al. to create patterns of pure color perturbations [101]. Both two approaches determine the optimal light parameters through query-based methods. Natural phenomena are also utilized to design optical perturbations. Zhong et al. [102] employed the shadow effect, as showcased in Fig. 13 (c), to design the loss function based on the picture luminance and used PSO to calculate shadow edge positions. Hu et al. [103] adopted mirrors to reflect query-based adversarial catoptric light on the target position of a traffic sign, as depicted in Fig. 13 (d). Furthermore, using a specific optical device to engage with the victim vehicle may result in traffic sign misclassification. Hu et al. [104] put a zoom lens before the camera, dynamically adjusting magnification and continuously querying the prediction model until the original label achieved the lowest scores.

Attack trajectory prediction by optical methods. Projectors can also interfere with other essential visual tasks, including object tracking. Muller et al. [105] suggested employing projections to mislocate the target bounding box, as depicted in Fig. 14. They initially analyzed the point cloud to determine the area of interest for the camera, such as a car or a traffic sign. Perturbations were then created in that region. Afterward, a siamese model with accessible gradients is utilized to train these perturbations.

Attack trajectory prediction by trajectory. An attacker possesses the capability to manipulate the driving trajectory of their vehicle and cause the victim vehicle to inaccurately

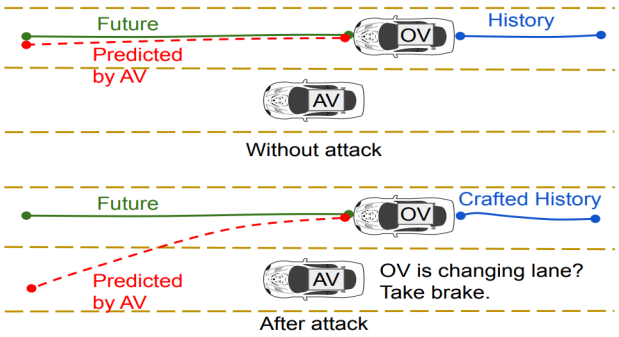


Fig. 15. Attack trajectory prediction by trajectory (image credit: [107]).

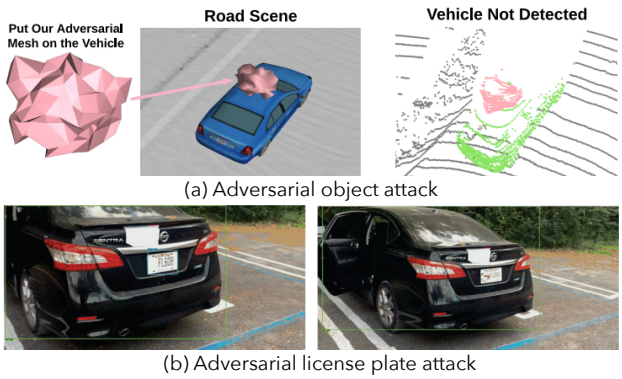


Fig. 16. Attack vehicle detection by adversarial objects (image credit: [11], [109]).

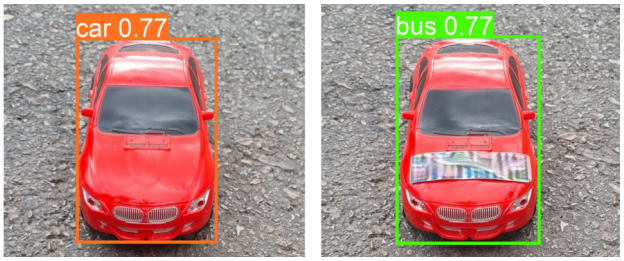


Fig. 17. Attack vehicle detection by patch (image credit: [110]).

forecast the attacking vehicle's future route (as illustrated in Fig. 15). The victim vehicle turns to avoid an anticipated collision, resulting in a real accident. Cao et al. [106] utilized publicly available trajectory prediction models and the PGD algorithm to generate transferable adversarial trajectories. While Zhang et al. [107] employed the PSO algorithm to search adversarial trajectories. Additional constraints were added to the optimization problem to ensure that the trajectories follow traffic regulations. The trajectory attack effectively led to a collision involving the victim vehicle, as demonstrated in the LGSVL simulator.

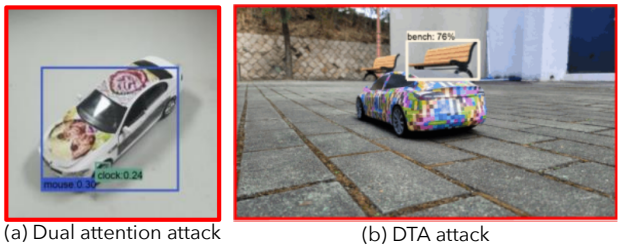
Attack LiDAR prediction by trajectory. Data collection may be delayed during LiDAR rotation in the event of a moving vehicle. The delay may lead to distortions in the point cloud. ADSs perform motion compensation before creating the point cloud to address the problem. Li et al. [108] developed a trajectory attack that exploits vulnerabilities in the motion compensation model. The attack involved introducing perturbations to the vehicle trajectory, i.e., allowing the victim vehicle to travel along the adversarial trajectory so that safety objects were either missed or incorrectly detected. The researchers simulated motion compensation models and created a differentiable function between the point cloud and the vehicle trajectory. The function was then used to construct the loss function and the transferable perturbation was trained through the PGD method.

Attack vehicle detection by adversarial objects. Adversarial objects can attack both the camera-based and LiDAR detection tasks in MSF. So they can indeed effectively attack each type of detection independently. Tu et al. [11] concentrated on attacking the LiDAR-based vehicle detection task. The researchers connected the point cloud of the adversarial mesh to the top of the vehicle. Then they used a genetic algorithm for optimization. The adversarial object successfully made the carrier invisible under LiDAR-based vehicle detection (as shown in Fig. 16 (a)). Yang et al. [109] engineered an adversarial license plate that was misidentified as part of the background in camera-based detection by formulating a tailored loss function. They iteratively optimized the adversarial object using a gradient-based technique. Various vehicles and physical contexts were substituted to improve the attack transferability during the training phase. Fig. 16 (b) shows that the adversarial license plate, made of aluminum, successfully disrupted license plate recognition.

Attack vehicle detection by patch. Shapira et al. [110] developed an innovative adversarial patch, as shown in Fig. 17, which can be placed on the front of a vehicle to trick the detector and cause misclassification. The researchers carried out patch training with a video stream of moving vehicles. They applied the patch to all vehicles that had complete bounding boxes within the video to improve the patch's universality. Their objective was to decrease the confidence score for identifying cars while increasing it for buses by designing a loss function. To assess the effectiveness of the attack they printed these patches out and attached them to a toy car's hood. The patch carrier's movements were recorded and subsequently evaluated with a pre-trained model.

Attack vehicle detection by camouflage. Placing a large camouflage on the target vehicle can deceive the vision-based vehicle detector on the victim. Wang et al. [111] designed a loss that removes the model's attention away from the target vehicle. The researchers integrated adversarial perturbations with common visual patterns in an effort to make the attack more stealthy. As shown in Fig. 18(a), placing the camouflage on a large, flat area like the roof or front hood of the target vehicle can result in this vehicle being misclassified by the ADS. Wu et al. [112] devised a loss function with the aim of increasing the occurrence of vehicle misclassifications. A discrete search algorithm is used to optimize the perturbation. The physical deployment strategy involves enlarging and repeatedly applying the generated square perturbation to the target vehicle.

Wang et al. [113] employed a differentiable neural rendering network approach to deal with the transition from 2D plane to 3D space. They improved the camouflage robustness by



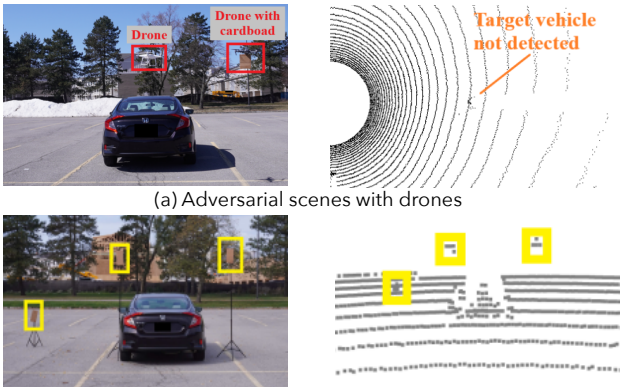
(a) Dual attention attack (b) DTA attack

Fig. 18. Attack vehicle detection by camouflage (image credit: [111], [114]).

iteratively updating the surroundings of the target vehicle using a transformation function. However, neural rendering typically focuses on the foreground image from the camera. Suryanto et al. [114] attempted to evade detection from various viewpoints by covering camouflage over the target vehicle. They developed a differential transformation network (DTN) that was capable of learning scene attributes, including material properties, lighting influences, shadows, etc. The optimization process relied on leveraging DTN gradients. Implementation results, as seen in Fig. 18 (b), underscored the effectiveness of their approach. These camouflage attacks were simulated within the CARLA and were also applied to toy cars for real-world experimentation.

Attack vehicle detection by adversarial scene. Zhu et al. [12], [115] conducted their study on scene attacks against vehicle detection and shared their findings in two related papers. The first paper discussed designing an attack strategy that uses a drone hovering at an adversarial position to cause the front vehicle to be undetected (as depicted in Fig. 19 (a)) [12]. The main challenge of the approach was to minimize the number of these positions while maintaining the confidence level of the targeted bounding box below the detection threshold. A heuristic-based approach was proposed to identify the locations. The second paper was concerned with selecting objects commonly found on real roads and placing them in adversarial locations (as shown in Fig. 19 (b)). Their objective was to manipulate the semantic features of the adversarial point cloud to resemble those in specific dangerous scenes, i.e., targeted attacks [115]. To achieve a query-based attack the authors employed the differential evolution (DE) algorithm [116]. Both articles involved installing a LiDAR system on top of a sedan to simulate real-world data collection.

Attack vehicle detection by sound waves. The concept of sound waves causing vibrations in devices was leveraged by Ji et al. [117] to design adversarial acoustic signals. When these signals are transmitted, the cameras capture blurred images. These images can trick object detection on ADS by making the targeted vehicle disappear (as shown in Fig. 20). A model for fuzzy patterns was developed so as to make these manipulations applicable to real signals. The model considered the relationship between the generator and the output of the shaking sensor. The main target of the optimization problem was to minimize the confidence level associated with the bounding box. The target was achieved by using Bayesian optimization, where a surrogate model was established to approximate the black-box detector. Subsequently, the gradients and outputs from the surrogate model were used to identify the possible solution.



(a) Adversarial scenes with drones

(b) Attack scene with natural objects

Fig. 19. Attack vehicle detection by scene (image credit: [12], [115]).

(a) Sound waves attack (Samsung S20 in Audi Q3)

(b) Detection results before and after the attack

Fig. 20. Attack vehicle detection by sound waves (image credit: [117]).

B. Sensor-based attack

A more profound understanding of the physics behind sensor operation is undoubtedly essential for sensor-based attacks. In camera imaging, light is focused onto an image sensor through a lens. The photosensitive elements on the sensor convert this light to an electrical signal, which is subsequently converted to a digital image by processing circuitry. Researchers have discovered potential dangers upon gaining a deeper understanding of this process, such as the light refraction between lenses to form flares (Section III-B2) and the ability of photosensitive elements to capture invisible light. LiDAR, mmW radar, and ultrasonic radar work on a similar principle with the difference in carriers. They all measure the time from the transmitted to the received signals to determine the distance to the target. So the idea behind their attacks is similar or emitting a fraudulent signal. Researchers use the phenomena of wave superposition, interference, and diffraction to manipulate the number, frequency, direction, and other physical factors of the returning waves.

Attack radar perception by mmWaves. The victim vehicle lacks the capability to determine whether the signal it receives is genuine or not, making the receiver vulnerable to malicious signals sent from an mmWave transmitter. Sun et al. [118]

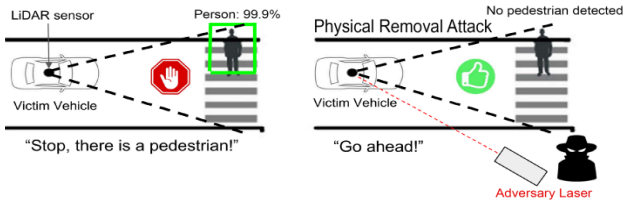
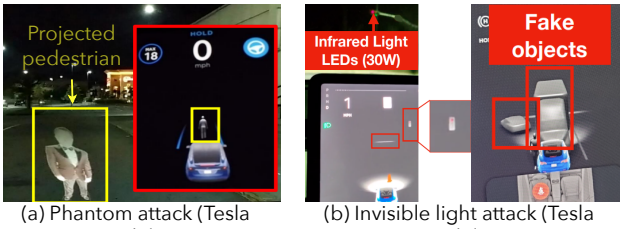
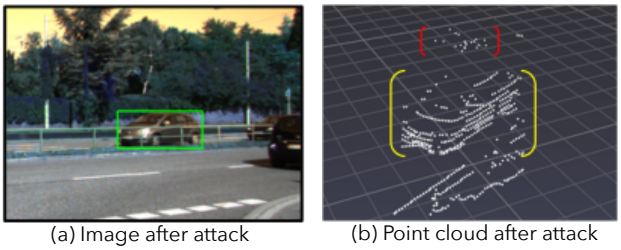


Fig. 21. Attack LiDAR perception by laser (image credit: [121]).



(a) Phantom attack (Tesla Model X) (b) Invisible light attack (Tesla Model 3)

Fig. 22. Attack camera perception by optical methods (image credit: [122], [123])



(a) Image after attack (b) Point cloud after attack

Fig. 23. Attack MSF perception by laser (image credit: [124])

recorded the signal reflected by an obstacle and subsequently replicated that signal with a delay, inducing the victim vehicle to believe there was another obstacle ahead. Moreover, the authors presented another method to manipulate an obstacle's perceived location. In an effort to faithfully mimic the position of the obstacle, they had to cancel out the reflected signal from the original obstacle. A series of experiments were conducted on the Lincoln MKZ to validate the effectiveness of the attack.

Attack LiDAR perception by optical methods. The imitated laser received by the ADS LiDAR receiver can cause the interpretation of a fabricated point cloud. This attack can disable LiDAR detection as shown in Fig. 21. In a related study of Jin et al. [119], they captured and recorded an obstacle's point cloud using a radar similar to the one in the victim vehicle. A replay of the received signals was then performed in order to evaluate the performance of the VLP-16 point cloud under various laser emitters. In this way, they were able to find the optimal emitter parameters. Furthermore, vehicles far away or obscured by another can still be correctly classified even if they possess a few points. Sun et al. [120] exploited this susceptibility by creating sparse points using a laser emitter ahead of the victim vehicle. Consequently, the victim's ADS detected a non-existent vehicle in front of it, achieving a deceptive attack with a reduced number of point clouds.

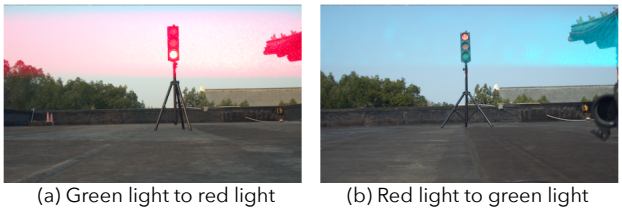
A laser beam may strike multiple objects along its propagation path, resulting in echoes that are too complex to compute. This is why most LiDAR systems are configured to detect the strongest echoes. Cao et al. [121] proposed an object removal attack exploiting this rule, where the attacker transmits the strongest echo near the sensor at the appropriate time. The behavior results in the LiDAR receiver ignoring true echoes, clearing the point cloud in the obstacle area. The authors simulated their object removal technique in LGSVL to demonstrate the attack's impact.

Attack camera perception by optical methods. Sometimes our visual perception may not always accurately reflect reality. Nassi et al. [122] proposed an attack involving projectors, drone projections, and digital billboards to display phantoms of real-world objects. This approach has the potential to convince the autopilot system that the phantom is real (as demonstrated in Fig. 22 (a)). In order to determine the optimal parameters of the phantom, the authors conducted experiments with phantoms of varying sizes and intensities on a vehicle equipped with an advanced driver assistance system (ADAS), the Mobileye 630 PRO. Wang et al. [123] proposed to disrupt camera perception by emitting imperceptible infrared light. Their experiment involved using light within the range of $780nm - 850nm$ to generate traffic signs, obstacles, and can even blind the camera (as shown in Fig. 22 (b)). Both of these attacks are carried out on a target Tesla in a manner that does not alert the driver.

Attack MSF perception by optical methods. Hallyburton et al. [124] introduced a method known as the Frustum attack. The attack takes advantage of the fact that cameras capture 2D representations of AD scenes lacking spatial details. The attacker manipulated the laser transmitter to modify the point cloud received by the victim's LiDAR while ensuring semantic coherence between the image and point cloud data. Essentially the obstacle remains visually present but its position information provided by the point cloud is altered. The attack result is illustrated in Fig. 23 where the injected fraudulent points are in the red box in (b). Their presence causes the MSF to detect the obstacle (green in (a)) at an incorrect location (red in (b)) instead of the correct one (yellow in (b)).

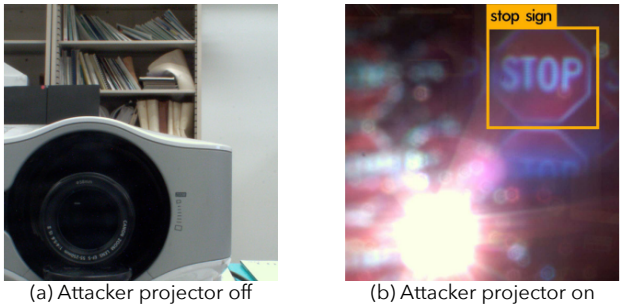
Attack camera-based depth estimation by optical methods. Zhou et al. [125] developed a method for carrying out depth estimation attacks specifically targeting stereo cameras. Normally these cameras rely on matching a pair of images from the left and right sides to gather information about their surroundings. The researchers exploited the flare effect discussed in SectionIII-B2 by using beams or spheres of light to bring glare to the images. The attack spoofed depth estimation by creating a false stereo correspondence between the left and right images with these glares. As proof of attack efficiency, they used two projectors to irradiate the DJI drone, resulting in inaccurate measurements of distance.

Attack traffic light recognition by optical methods. Yan et al. [126] conducted a study in which a significant security vulnerability in the CMOS camera's rolling shutter function was discovered. Researchers overlaid colors on traffic signs with laser injection techniques to cause the red and green signals to be misidentified (as shown in Fig. 24). The authors start by developing a model to simulate laser interference



(a) Green light to red light (b) Red light to green light

Fig. 24. Attack traffic light recognition by laser (image credit: [126]).



(a) Attacker projector off (b) Attacker projector on

Fig. 25. Attack traffic sign recognition by projector (image credit: [67]).

accurately and then construct an optimization problem based on the model. A grid search method was used to determine the optimal parameters for a laser device that reverses traffic light color judgments.

Attack traffic sign recognition by optical methods. Man et al. [67] exploited the lens flare effect, a phenomenon where images of traffic signs projected onto a camera lens could be clearly captured. In cases where the attacker possessed an identical camera to that of the victim vehicle, a model of the camera was simulated. This model can predict the ghost image's location, resolution, and color based on parameters such as light source intensity and camera position. It was possible for authors to find the parameters where the attack is most effective through the model. Subsequently, as shown in Fig. 25, they used a projector to present the ideal attack ghosting to the victim camera.

C. Defense challenge

In the arms race between defenses and attacks, defenses are more passive as they always respond to attacks. Fig. 26 shows a timeline of the attacks investigated in this study and the defenses during the same period. If the attacks only target one sensor, ADSs can significantly reduce the rate of contamination in collected data by using redundant sensors or MSF [127]. On the other hand, MSF attacks [14] that disrupt the stability of heterogeneous information are on the rise while effective defenses against these attacks are still in development. Currently, the most effective method for traditional adversarial attacks is adversarial training [128], [129]. However, retraining models is indeed expensive and time-consuming, making it challenging to update them promptly in response to attacks.

There are also data augmentation-based methods [130], preprocessing-based methods [131], and approaches with additional networks [132], [133]. For example, Vitale et al. [133] allowed anti-hacking devices to integrate AI/ML-based AE detectors in their CARAMEL project. Despite these efforts, these defenses can be evaded when attackers successfully convert the structure to a white-box [134].

A single sensor can handle different tasks, and some of them corroborate with each other's results. You et al. [135] detected a LiDAR perception attack by comparing the results from an object motion predictor with those from an object detector. Defenders are actively exploring the operational principles of sensors. Lou et al. [136] randomized the pulse period of the ultrasonic radar to make it difficult for attackers to predict the operating pattern of sensor.

Besides eliminating perturbations, defenders also look for ways to regenerate the original image. Wang et al. [137] combined generative models with techniques like compressive sensing and adversarial training to propose their defense. Moreover, taking advantage of the strong performance of diffusion models in recent years, many strategies have been applied in defenses with the help of their understanding and generating abilities [138], [139]. Nevertheless, training these models requires a long time and significant computing power, which presents a challenge for their practical implementation on intelligent vehicles.

In terms of the equipment, many researchers have analyzed the physics of an attack and attempted to filter it out when it reaches the sensor using special materials, such as polarizer [93], [140]. High innovation in sensor technology has been apparent over the past few years. However, rising costs have left most manufacturers with a choice of value over performance.

V. Perspectives and Future Directions

The purpose of the survey is to acquire a comprehensive understanding of the most recent physical attacks. A multi-tag approach was discussed to clearly define each attack through investigation. The tag combination can explore the principles behind physical attack design and implementation, and identify physical threats to safe driving in the real world. A wide range of functions, such as sensors and networks, have been seamlessly integrated into ADSs, and continually improving. Due to continuous upgrades of their internal firewalls, conducting internal ADS attacks have become increasingly challenging. The flexibility and versatility of the physical attacks often yield surprising results.

As ADS technology evolves, the developing trends and requirements of physical attacks on ADS undergo significant shifts. Following our analysis, we draw five key conclusions: (1) Attack targets range from individual DNNs to the entire system with sensors and networks. Moreover, a combination of attack forms is potentially used. (2) Laser-based attacks have become increasingly popular recently. These attacks are inexpensive, flexible, and stealthy in execution, while also breaking traditional constraints at the pixel level. (3) Evaluation methods are diverse. Some studies use only the output of perception networks for assessment, while other attackers prefer visualizing their attack consequences on actual vehicles or simulators. (4) There is an absence of risk analysis that determines the level of danger associated with physical attacks. The traditional attack success rate, ASR (%) =

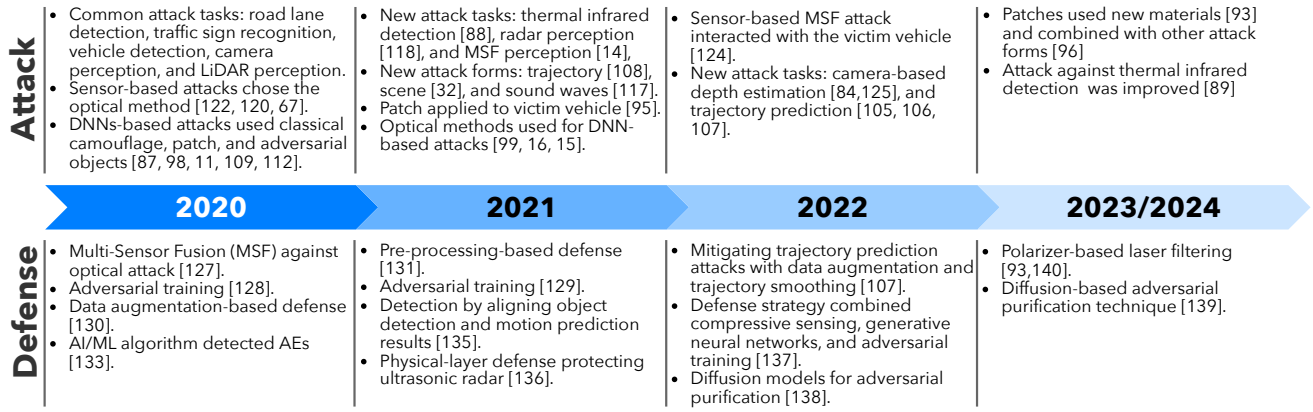


Fig. 26. Timeline of physical attack from 2020 - 2024 and defenses in the same period.

$\frac{\text{Number of Successful Samples}}{\text{Total Number of Samples}} \times 100$, is insufficient to express the complex physical consequences. (5) Attackers or defenders need clear threat frameworks to assist in decision-making. Therefore, we outline future directions corresponding to the above five points:

- Designing attacks that deceive the perception of MSF with the consideration of the entire ADS, exploring combinations of attack forms.
- Developing detection and defense against physical attacks, specifically laser-based attacks as demonstrated in our Laser Shield [140].
- Establishing an evaluation methodology and metrics to assess attack effectiveness and success.
- Proposing a risk-impact assessment, considering real-life consequences and driver or road safety.
- Using attack modeling paradigms like attack trees or graphs etc., to cover all potential vulnerabilities/sploits and threats for each part of the ADS perception system.

VI. Conclusion

The security of current ADS considering physical attacks has become vital because of the increasing integration of ADS into our daily lives. The evolution of physical attacks against ADS perception is comprehensively analyzed and a four-dimensional classification for categorizing them is proposed in the survey. According to our analysis, the attack strategy has evolved from targeting isolated DNNs to the entire ADS ecosystem, from creating pixel-level digital attacks to physical attacks that exploit natural phenomena, as well as from applying static stickers to dynamic light displays. The shift is because attackers have an advantage in real-world attack setups since defenders cannot control real-world scenarios in real time. The suggested directions for future research are the development of MSF attacks with multiple-form attacks, physical attack detection and defense, a comprehensive paradigm for attack evaluation, risk analysis, and attack modeling.

References

[1] L. Chen, Y. Li, C. Huang, B. Li, Y. Xing, D. Tian, L. Li, Z. Hu, X. Na, Z. Li, S. Teng, C. Lv, J. Wang, D. Cao, N. Zheng, and F.-Y. Wang, "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046–1056, 2023.

[2] P. S. Chib and P. Singh, "Recent advancements in end-to-end autonomous driving using deep learning: A survey," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 103–118, 2024.

[3] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[5] H. Qiu, T. Dong, T. Zhang, J. Lu, G. Memmi, and M. Qiu, "Adversarial attacks against network intrusion detection in iot systems," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10 327–10 335, 2020.

[6] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020.

[7] A. Hamdi, S. Rojas, A. Thabet, and B. Ghanem, "Advpc: Transferable adversarial perturbations on 3d point clouds," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 241–257.

[8] X. Wu, S. Ma, C. Shen, C. Lin, Q. Wang, Q. Li, and Y. Rao, "KENKU: Towards efficient and stealthy black-box adversarial attacks against ASR systems," in *32nd USENIX Security Symposium*, 2023, pp. 247–264.

[9] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.

[10] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293.

[11] J. Tu, M. Ren, S. Manivasagam, M. Liang, B. Yang, R. Du, F. Cheng, and R. Urtasun, "Physically realizable adversarial examples for lidar object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 716–13 725.

[12] Y. Zhu, C. Miao, T. Zheng, F. Hajiaghajani, L. Su, and C. Qiao, "Can we use arbitrary objects to attack lidar perception in autonomous driving?" in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 1945–1960.

[13] T. Sato, J. Shen, N. Wang, Y. Jia, X. Lin, and Q. A. Chen, "Dirty road can attack: Security of deep learning based automated lane centering under physical-world attack," in *30th USENIX Security Symposium*, 2021, pp. 3309–3326.

[14] Y. Cao, N. Wang, C. Xiao, D. Yang, J. Fang, R. Yang, Q. A. Chen, M. Liu, and B. Li, "Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under

physical-world attacks," in *2021 IEEE Symposium on Security and Privacy*. IEEE, 2021, pp. 176–194.

[15] R. Duan, X. Mao, A. K. Qin, Y. Chen, S. Ye, Y. He, and Y. Yang, "Adversarial laser beam: Effective physical-world attack to dnns in a blink," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 062–16 071.

[16] A. Gnanasambandam, A. M. Sherman, and S. H. Chan, "Optical adversarial attack," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 92–101.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[18] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger. in proceedings of the ieee conference on computer vision and pattern recognition (pp. 7263-7271)," 2017.

[19] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, J. Fang, K. Michael, D. Montes, J. Nadar, P. Skalski *et al.*, "ultralytics/yolov5: v6. 1-tensorrt, tensorflow edge tpu and openvino export and inference," *Zenodo*, 2022.

[20] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[25] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[26] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 770–779.

[27] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7652–7660.

[28] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8971–8980.

[29] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 101–117.

[30] X. Li, X. Ying, and M. C. Chuah, "Grip++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving," *arXiv preprint arXiv:1907.07792*, 2019.

[31] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 683–700.

[32] W. Ding, H. Lin, B. Li, K. J. Eun, and D. Zhao, "Semantically adversarial driving scenario generation with explicit knowledge integration," *arXiv e-prints*, pp. arXiv–2106, 2021.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[36] B. Wu, A. Wan, X. Yue, and K. Keutzer, "Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1887–1893.

[37] H. Zhou, X. Zhu, X. Song, Y. Ma, Z. Wang, H. Li, and D. Lin, "Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation," *arXiv preprint arXiv:2008.01550*, 2020.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[42] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: a multi-class classification competition," in *The 2011 international joint conference on neural networks*. IEEE, 2011, pp. 1453–1460.

[43] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE transactions on intelligent transportation systems*, vol. 13, no. 4, pp. 1484–1497, 2012.

[44] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2110–2118.

[45] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[48] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[49] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[50] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[51] P. A. Sava, J.-P. Schulze, P. Sperl, and K. Böttinger, "Assessing the impact of transformations on physical adversarial attacks," in *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*, 2022, pp. 79–90.

[52] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293.

[53] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, "Robust physical-world attacks on machine learning models," *arXiv preprint arXiv:1707.08945*, vol. 2, no. 3, p. 4, 2017.

[54] M. Xue, C. Yuan, C. He, J. Wang, and W. Liu, "Naturalae: Natural and robust physical adversarial examples for object detectors," *Journal of Information Security and Applications*, vol. 57, p. 102694, 2021.

[55] A. Qayyum, M. Usama, J. Qadir, and A. Al-Fuqaha, "Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 998–1026, 2020.

[56] Y. Deng, T. Zhang, G. Lou, X. Zheng, J. Jin, and Q.-L. Han, "Deep learning-based autonomous driving systems: A survey of attacks and defenses," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 7897–7912, 2021.

[57] F. Woitschek and G. Schneider, "Physical adversarial attacks on deep neural networks for traffic sign recognition: A feasibility study," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 481–487.

[58] K. T. Y. Mahima, M. Ayoob, and G. Poravi, "Adversarial attacks and defense technologies on autonomous vehicles: A review," *Appl. Comput. Syst.*, vol. 26, no. 2, p. 96–106, dec 2021.

[59] C. Gao, G. Wang, W. Shi, Z. Wang, and Y. Chen, "Autonomous driving security: State of the art and challenges," *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7572–7595, 2021.

[60] X. Wei, B. Pu, J. Lu, and B. Wu, "Physically adversarial attacks and defenses in computer vision: A survey," *arXiv preprint arXiv:2211.01671*, 2022.

[61] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[62] B. Sanyal, R. K. Mohapatra, and R. Dash, "Traffic sign recognition: A survey," in *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*. IEEE, 2020, pp. 1–6.

[63] N. Suganuma and K. Yoneda, "Current status and issues of traffic light recognition technology in autonomous driving system," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 105, no. 5, pp. 763–769, 2022.

[64] S. Sheng, N. Formosa, M. Hossain, and M. Quddus, "Advancements in lane marking detection: An extensive evaluation of current methods and future research direction," *IEEE Transactions on Intelligent Vehicles*, pp. 1–12, 2024.

[65] Z. Wang, J. Zhan, C. Duan, X. Guan, P. Lu, and K. Yang, "A review of vehicle detection techniques for intelligent vehicles," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[66] J. Miao, K. Jiang, T. Wen, Y. Wang, P. Jia, B. Wijaya, X. Zhao, Q. Cheng, Z. Xiao, J. Huang, Z. Zhong, and D. Yang, "A survey on monocular re-localization: From the perspective of scene map representation," *IEEE Transactions on Intelligent Vehicles*, pp. 1–33, 2024.

[67] Y. Man, M. Li, and R. Gerdes, "GhostImage: Remote perception attacks against camera-based image classification systems," in *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*, 2020, pp. 317–332.

[68] N. H. H. Aung, P. Sangwongngam, R. Jintamethasawat, S. Shah, and L. Wuttisittikulkij, "A review of lidar-based 3d object detection via deep learning approaches towards robust connected and autonomous vehicles," *IEEE Transactions on Intelligent Vehicles*, pp. 1–23, 2024.

[69] A. Venon, Y. Dupuis, P. Vasseur, and P. Merriaux, "Millimeter wave fmcw radars for perception, recognition and localization in automotive applications: A survey," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 533–555, 2022.

[70] A. Rangesh and M. M. Trivedi, "No blind spots: Full-surround multi-object tracking for autonomous vehicles using cameras and lidars," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 4, pp. 588–599, 2019.

[71] S. Yao, R. Guan, X. Huang, Z. Li, X. Sha, Y. Yue, E. G. Lim, H. Seo, K. L. Man, X. Zhu, and Y. Yue, "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 2094–2128, 2024.

[72] Y. He, B. Deng, H. Wang, L. Cheng, K. Zhou, S. Cai, and F. Ciampa, "Infrared machine vision and infrared thermography with deep learning: A review," *Infrared physics & technology*, vol. 116, p. 103754, 2021.

[73] L. Jin, L. Liu, X. Wang, M. Shang, and F.-Y. Wang, "Physical-informed neural network for mpc-based trajectory tracking of vehicles with noise considered," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 3, pp. 4493–4503, 2024.

[74] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, "A survey on trajectory-prediction methods for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 652–674, 2022.

[75] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.

[76] S. Chen, Z. He, C. Sun, J. Yang, and X. Huang, "Universal adversarial attack on attention and the resulting dataset damagenet," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[77] L. Chi, M. Msahli, G. Memmi, and H. Qiu, "Public-attention-based adversarial attack on traffic sign recognition," in *2023 IEEE 20th Consumer Communications & Networking Conference*. IEEE, 2023, pp. 740–745.

[78] P. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

[79] J. Chen, M. Su, S. Shen, H. Xiong, and H. Zheng, "Poba-ga: Perturbation optimized black-box adversarial attacks via genetic algorithm," *Computers & Security*, vol. 85, pp. 89–106, 2019.

[80] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. ieee, 1995, pp. 1942–1948.

[81] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: a big comparison for nas," *arXiv preprint arXiv:1912.06059*, 2019.

[82] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[83] K. Yang, T. Tsai, H. Yu, M. Panoff, T.-Y. Ho, and Y. Jin, "Robust roadside physical adversarial attack against deep learning in lidar perception modules," in *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 2021, pp. 349–362.

[84] Z. Cheng, J. Liang, H. Choi, G. Tao, Z. Cao, D. Liu, and X. Zhang, "Physical attack on monocular depth estimation with optimal adversarial patches," in *European conference on computer vision*. Springer, 2022, pp. 514–532.

[85] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on scientific computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[86] P. Jing, Q. Tang, Y. Du, L. Xue, X. Luo, T. Wang, S. Nie, and S. Wu, "Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations," in *30th USENIX Security Symposium*, 2021, pp. 3237–3254.

[87] A. Boloor, K. Garimella, X. He, C. Gill, Y. Vorobeychik, and X. Zhang, "Attacking vision-based perception in end-to-end autonomous driving models," *Journal of Systems Architecture*, vol. 110, p. 101766, 2020.

[88] X. Zhu, X. Li, J. Li, Z. Wang, and X. Hu, "Fooling thermal infrared pedestrian detectors in real world using small bulbs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3616–3624.

[89] H. Wei, Z. Wang, X. Jia, Y. Zheng, H. Tang, S. Satoh, and Z. Wang, "Hotcold block: Fooling thermal infrared detectors with a novel wearable design," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 12, 2023, pp. 15 233–15 241.

[90] X. Wei, Y. Guo, J. Yu, and B. Zhang, "Simultaneously optimizing perturbations and positions for black-box adversarial patch attacks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[91] X. Wei, Y. Guo, and J. Yu, "Adversarial sticker: A stealthy attack method in the physical world," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[92] L. Giulivi, M. Jere, L. Rossi, F. Koushanfar, G. Ciocarlie, B. Hitaj, and G. Boracchi, "Adversarial scratches: Deployable attacks to cnn classifiers," *Pattern Recognition*, vol. 133, p. 108985, 2023.

[93] G. Tsuruoka, T. Sato, Q. A. Chen, K. Nomoto, Y. Tanaka, R. Kobayashi, and T. Mori, "Wip: Adversarial retroreflective patches: A novel stealthy attack on traffic sign recognition at night," *Network and Distributed System Security (NDSS) Symposium*, 2024.

[94] C. Hu and W. Shi, "Adversarial color film: Effective physical-world attack to dnns," *arXiv preprint arXiv:2209.02430*, 2022.

[95] A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai, "The translucent patch: A physical and universal attack on object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 232–15 241.

[96] W. Zhu, X. Ji, Y. Cheng, S. Zhang, and W. Xu, "Tpatch: A triggered physical adversarial patch," in *32st USENIX Security Symposium*, 2023.

[97] W. Jia, Z. Lu, H. Zhang, Z. Liu, J. Wang, and G. Qu, "Fooling the eyes of autonomous vehicles: Robust physical adversarial examples against traffic sign recognition systems," *arXiv preprint arXiv:2201.06192*, 2022.

[98] X. Yang, W. Liu, S. Zhang, W. Liu, and D. Tao, "Targeted attention attack on deep learning models in road sign recognition," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4980–4990, 2020.

[99] G. Lovisotto, H. Turner, I. Sluganovic, M. Strohmeier, and I. Martinovic, "SLAP: Improving physical adversarial examples with Short-Lived adversarial perturbations," in *30th USENIX Security Symposium*, 2021, pp. 1865–1882.

[100] S. Xie, H. Wang, Y. Kong, and Y. Hong, "Universal 3-dimensional perturbations for black-box attacks on video recognition systems," in *2022 IEEE Symposium on Security and Privacy*. IEEE, 2022, pp. 1390–1407.

[101] R. Hu, T. Rui, Y. Ouyang, J. Wang, Q. Jiang, and Y. Du, "Light attack: A physical world real-time attack against object classifiers," *IEEE Access*, 2022.

[102] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji, "Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 345–15 354.

[103] C. Hu, W. Shi, L. Tian, and W. Li, "Adversarial catoptric light: An effective, stealthy and robust physical-world attack to dnns," *IET Computer Vision*, 2024.

[104] C. Hu and W. Shi, "Adversarial zoom lens: A novel physical-world attack to dnns," *arXiv preprint arXiv:2206.12251*, 2022.

[105] R. Muller, Y. Man, Z. B. Celik, M. Li, and R. Gerdes, "Physical hijacking attacks against object trackers," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2309–2322.

[106] Y. Cao, C. Xiao, A. Anandkumar, D. Xu, and M. Pavone, "Advdo: Realistic adversarial attacks for trajectory prediction," in *European Conference on Computer Vision*. Springer, 2022, pp. 36–52.

[107] Q. Zhang, S. Hu, J. Sun, Q. A. Chen, and Z. M. Mao, "On adversarial robustness of trajectory prediction for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 159–15 168.

[108] Y. Li, C. Wen, F. Juefei-Xu, and C. Feng, "Fooling lidar perception via adversarial trajectory perturbation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7898–7907.

[109] K. Yang, T. Tsai, H. Yu, T.-Y. Ho, and Y. Jin, "Beyond digital domain: Fooling deep learning based recognition system in physical world," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 1088–1095.

[110] A. Shapira, R. Bitton, D. Avraham, A. Zolfi, Y. Elovici, and A. Shabtai, "Attacking object detector using a universal targeted label-switch patch," *arXiv preprint arXiv:2211.08859*, 2022.

[111] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu, "Dual attention suppression attack: Generate adversarial camouflage in physical world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8565–8574.

[112] T. Wu, X. Ning, W. Li, R. Huang, H. Yang, and Y. Wang, "Physical adversarial attack on vehicle detector in the carla simulator," *arXiv preprint arXiv:2007.16118*, 2020.

[113] D. Wang, T. Jiang, J. Sun, W. Zhou, Z. Gong, X. Zhang, W. Yao, and X. Chen, "Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2414–2422.

[114] N. Suryanto, Y. Kim, H. Kang, H. T. Larasati, Y. Yun, T.-T.-H. Le, H. Yang, S.-Y. Oh, and H. Kim, "Dta: Physical camouflage attacks using differentiable transformation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 305–15 314.

[115] Y. Zhu, C. Miao, F. Hajiaghajani, M. Huai, L. Su, and C. Qiao, "Adversarial attacks against lidar semantic segmentation in autonomous driving," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 2021, pp. 329–342.

[116] R. Storn, "On the usage of differential evolution for function optimization," in *Proceedings of north american fuzzy information processing*. Ieee, 1996, pp. 519–523.

[117] X. Ji, Y. Cheng, Y. Zhang, K. Wang, C. Yan, W. Xu, and K. Fu, "Poltergeist: Acoustic adversarial machine learning against cameras and computer vision," in *2021 IEEE Symposium on Security and Privacy*. IEEE, 2021, pp. 160–175.

[118] Z. Sun, S. Balakrishnan, L. Su, A. Bhuyan, P. Wang, and C. Qiao, "Who is in control? practical physical layer attack and defense for mmwave-based sensing in autonomous vehicles," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3199–3214, 2021.

[119] Z. Jin, J. Xiaoyu, Y. Cheng, B. Yang, C. Yan, and W. Xu, "Pla-lidar: Physical laser attacks against lidar-based 3d object detection in autonomous vehicle," in *2023 IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2022, pp. 710–727.

[120] J. Sun, Y. Cao, Q. A. Chen, and Z. M. Mao, "Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures," in *29th USENIX Security Symposium*, 2020, pp. 877–894.

[121] Y. Cao, S. H. Bhupathiraju, P. Naghavi, T. Sugawara, Z. M. Mao, and S. Rampazzi, "You can't see me: Physical removal attacks on lidar-based autonomous vehicles driving frameworks," in *32nd USENIX Security Symposium*, 2023, pp. 2993–3010.

[122] B. Nassi, Y. Mirsky, D. Nassi, R. Ben-Netanel, O. Drokin, and Y. Elovici, "Phantom of the adas: Securing advanced driver-assistance systems from split-second phantom attacks," in *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, 2020, pp. 293–308.

[123] W. Wang, Y. Yao, X. Liu, X. Li, P. Hao, and T. Zhu, "I can see the light: Attacks on autonomous vehicles using invisible lights," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 1930–1944.

[124] R. S. Hallyburton, Y. Liu, Y. Cao, Z. M. Mao, and M. Pajic, "Security analysis of camera-lidar fusion against black-box attacks on autonomous vehicles," in *31st USENIX Security Symposium*, 2022, pp. 1903–1920.

[125] C. Zhou, Q. Yan, Y. Shi, and L. Sun, "DoubleStar: Long-range attack towards depth estimation based obstacle avoidance in autonomous systems," in *31st USENIX Security Symposium*, 2022, pp. 1885–1902.

[126] C. Yan, Z. Xu, Z. Yin, S. Mangard, X. Ji, W. Xu, K. Zhao, Y. Zhou, T. Wang, G. Gu *et al.*, "Rolling colors: Adversarial laser exploits against traffic light recognition," in *31st USENIX Security Symposium*, 2022, pp. 1957–1974.

[127] J. Zhang, Y. Zhang, K. Lu, J. Wang, K. Wu, X. Jia, and B. Liu, "Detecting and identifying optical signal attacks on autonomous driving systems," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 1140–1153, 2020.

[128] Q. Sun, A. A. Rao, X. Yao, B. Yu, and S. Hu, "Counteracting adversarial attacks in autonomous driving," in *Proceedings of the 39th International Conference on Computer-Aided Design*, 2020, pp. 1–7.

[129] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," *arXiv preprint arXiv:2102.01356*, 2021.

[130] Y. Zeng, H. Qiu, G. Memmi, and M. Qiu, "A data augmentation-based defense method against adversarial attacks in neural networks," in *Algorithms and Architectures for Parallel Processing: 20th International Conference, New York City, NY, USA, October 2–4, 2020, Proceedings, Part II 20*. Springer, 2020, pp. 274–289.

[131] H. Qiu, Y. Zeng, Q. Zheng, S. Guo, T. Zhang, and H. Li, "An efficient preprocessing-based approach to mitigate advanced adversarial attacks," *IEEE Transactions on Computers*, vol. 73, no. 3, pp. 645–655, 2021.

[132] C. Kyrkou, A. Papachristodoulou, A. Kloukiniotis, A. Papandreou, A. Lalos, K. Moustakas, and T. Theocharides, "Towards artificial-intelligence-based cybersecurity for robustifying automated driving systems against camera sensor attacks," in *2020 IEEE computer society annual symposium on VLSI*. IEEE, 2020, pp. 476–481.

[133] C. Vitale, N. Piperigkos, C. Laoudias, G. Ellinas, J. Casademont, P. S. Khodashenas, A. Kloukiniotis, A. S. Lalos, K. Moustakas, P. B. Lobato *et al.*, "The caramel project: a secure architecture for connected and autonomous vehicles," in *2020 European Conference on Networks and Communications*. IEEE, 2020, pp. 133–138.

[134] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 3–14.

[135] C. You, Z. Hau, and S. Demetriou, "Temporal consistency checks to detect lidar spoofing attacks on autonomous vehicle perception," in *Proceedings of the 1st Workshop on Security and Privacy for Mobile AI*, 2021, pp. 13–18.

[136] J. Lou, Q. Yan, Q. Hui, and H. Zeng, "Soundfence: Securing ultrasonic sensors in vehicles using physical-layer defense," in *2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking*. IEEE, 2021, pp. 1–9.

[137] J. Wang, W. Su, C. Luo, J. Chen, H. Song, and J. Li, "Csg: Classifier-aware defense strategy based on compressive sensing and generative networks for visual recognition in autonomous vehicle systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9543–9553, 2022.

[138] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," *arXiv preprint arXiv:2205.07460*, 2022.

[139] K. Song, H. Lai, Y. Pan, and J. Yin, "Mimicdiffusion: Purifying adversarial perturbation via mimicking clean diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 665–24 674.

[140] Q. Zhang, L. Chi, D. Wang, M. Msahli, G. Memmi, T. Zhang, C. Zhang, and H. Qiu, "Laser shield: a physical defense with polarizer against laser attacks on autonomous driving systems," in *2024 61th ACM/IEEE Design Automation Conference*, 2024.

[141] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong, and H. Foroosh, "Polarnet: An improved grid representation for online lidar point clouds semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9601–9610.

[142] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9297–9307.

[143] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8748–8757.

[144] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 529–10 538.

[145] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[146] J. Watson, M. Firman, G. Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in *2019 IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, 2019, pp. 2162–2171.

[147] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1164–1174.

[148] H. Schafer, E. Santana, A. Haden, and R. Biasini, "A commute in data: The comma2k19 dataset," *arXiv preprint arXiv:1812.05752*, 2018.

[149] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[150] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[151] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[152] R. Lyu, "Nanodet-plus: Super fast and high accuracy lightweight anchor-free object detection model," *Available: https://github.com/RangiLyu/nanodet*, 2021.

[153] L. Larsen, *Learning Microsoft Cognitive Services*. Packt Publishing Ltd, 2017.

[154] L. Huang, "Traffic sign recognition database," *Available: https://nlpr.ia.ac.cn/pal/trafficdata/recognition.html*, 2020.

[155] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[156] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[157] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

[158] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[159] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[160] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[161] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[162] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[163] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[164] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[165] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[166] M. Kristan, J. Matas, A. Leonardis, T. Vojíř, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2137–2155, 2016.

[167] R. Muller, "Drivetruth: Automated autonomous driving dataset generation for security applications," in *Workshop on Automotive and Autonomous Vehicle Security (AutoSec)*, 2022.

[168] N. Kamra, H. Zhu, D. K. Trivedi, M. Zhang, and Y. Liu, "Multi-agent trajectory prediction with fuzzy query attention," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 530–22 541, 2020.

[169] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 954–960.

[170] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9813–9823.

[171] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head r-cnn: In defense of two-stage object detector," *arXiv preprint arXiv:1711.07264*, 2017.

[172] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[173] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[174] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.

[175] X. Yan, C. Zheng, Z. Li, S. Wang, and S. Cui, "Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5589–5598.

[176] P. Hu, J. Ziglar, D. Held, and D. Ramanan, "What you see is what you get: Exploiting visibility for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 001–11 009.

[177] C. Ertler, J. Mislej, T. Ollmann, L. Porzi, G. Neuhold, and Y. Kuang, "The mapillary traffic sign dataset for detection and classification on a global scale," in *European Conference on Computer Vision*. Springer, 2020, pp. 68–84.

[178] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

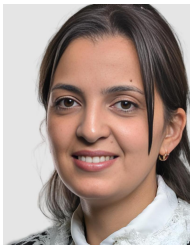
[179] G. Bradski, "The opencv library." *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, vol. 25, no. 11, pp. 120–123, 2000.

[180] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[181] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5410–5418.

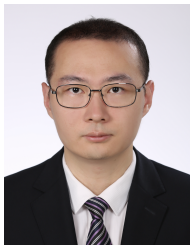
[182] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7134–7143.

Lijun Chi received the engineer degree in Mathematical Engineering from the Institut National des Sciences Appliquées de Rennes (INSA RENNES), Rennes, France, in 2019, and a dual M.S. degree in Applied Mathematics and Statistics from Université Rennes 2, Rennes, France, in 2019. She is currently working toward a Ph.D. degree in computer science with the Department of Networks and Computer Science (INFRES), Télécom Paris, Institut Polytechnique de Paris, France. Her research interests include computer vision, AI security, and ADS security.



Mounira Msahli received an M.Sc. in 2011 from Pierre and Marie Curie University, France, and a Ph.D on Security and Computer Science from Télécom Paris in 2015. She was a postdoc and research engineer at Télécom Paris from 2015 to 2018. She is currently an Associate Professor at the Network and Computer Science Department (INFRES), Télécom Paris. Her current research interests include the areas of vehicular network security and the use of IA for cybersecurity and misbehaving detection.

Qingjie Zhang received a B.E. and an M.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 2020 and 2023, and a dual M.S. degree from Télécom Paris, Institut Polytechnique de Paris, Paris, France, in 2023. He is currently a research assistant at the Institute for Network Sciences and Cyberspace, Tsinghua University. His research interests include trustworthy deep learning, large language models, and autonomous driving systems.



Han Qiu received a B.E. degree in 2011 from the Beijing University of Posts and Telecommunications, Beijing, China, an M.S. in 2013 from Télécom-ParisTech (Institute Eurecom), Biot, France, and a Ph.D. in 2017 in Computer Science from the Department of Networks and Computer Science, Télécom-Paris, Paris, France. He worked as a postdoc and research engineer at Télécom Paris and the LINCS Lab from 2017 to 2020. Currently, he is an Assistant Professor at the Institute for Network Sciences and Cyberspace, Tsinghua University, China. His research interests include AI security and edge computing.

Tianwei Zhang is an Assistant Professor at the School of Computer Science and Engineering at Nanyang Technological University. His research focuses on computer system security. He is particularly interested in security threats and defenses in machine learning systems, autonomous systems, computer architecture, and distributed systems. He received his Bachelor's degree at Peking University in 2011, and a Ph.D at Princeton University in 2017.

Gerard Memmi is professor and has been Head of the Networks and Computer Science Department at Télécom Paris since 2009. He has been a member of the executive board of the IRT SystemX since 2012. Before joining Télécom Paris, Gerard Memmi held various executive positions in American start-ups. He succeeded in delivering the industry's best-in-class equivalency checker used to verify electronic design and focused on improving its architecture and performances. While founding and developing the Applied Research Laboratory for Groupe Bull in the US, he was appointed Principal Investigator for a DARPA grant on Collaborative Software. He has over 100 publications including patents, co-authored a book, and delivered several key note presentations at international conferences. He holds a "these d'Etat" in Computer Science from the Université Pierre et Marie Curie, Paris. Today, his main research areas of interest are data protection and privacy, energy profiling of software programs, and formal verification of distributed systems.

Meikang Qiu received BE and ME degrees from Shanghai Jiao Tong University and received a Ph.D. in Computer Science from the University of Texas, Dallas. Currently, he is a full professor and director of AI enhanced Cyber Security Lab of Augusta University, USA. He is an ACM Distinguished Member. He is also the Highly Cited Researcher in 2021 from Web of Science and IEEE Distinguished Visitor in 2021-2023 as well as Chair of IEEE Smart Computing Technical Committee. Presently, his Google scholar citation is 22200+ and H-index 102.

APPENDIX

We list the black-box victims of these articles, namely the target system, DNNs, or sensors in Table III. One of the attacks with common stickers [91] proved its transfer capability only in face recognition, so we have filled in Table III with the target white-box traffic sign recognition model. The sensor-based attack can also be verified by the output of the perception model. Therefore, in the black-box victim column these attacks often have both sensors and models.

TABLE III
TAG COMBINATION FOR ATTACK METHODS.

Attack scene	Main factors of attack generation	Attack tasks	Attack forms	Papers	Black-box victim	Dataset for black-box attack generation and testing
Environment	DNNs	LiDAR perception	Scene	[32]	PointNet++ [35], SqueezeSegV3 [36], PolarSeg [141], Cylinder3D [37]	Semantic KITTI [142], Argoverse [143], scene in CARLA
			Adversarial objects	[83]	PointPillar [25], PV-RCNN [144]	KITTI [48]
		MSF perception		[14]	Baidu Apollo, Autoware (YOLOv3 [20])	COCO [47], KITTI [48]
		Camera-based depth estimation	Patch/Sticker	[84]	Monodepth2 [145], Depthhints [146], Manydepth [147]	KITTI [48], real photos
		Road lane detection		[86]	Tesla Model S75, with the Autopilot 2.5	Real photos
				[13]	Openpilot v0.7.0, v0.6.6, v0.5.9	Comma2k19 [148]
				[87]	CARLA simulator [50]	Scene in CARLA
		Thermal infrared detection		[88]	Cascade-RCNN [149], RetinaNet [150]	FLIR ADAS
				[89]	YOLOv3 [20], Faster-RCNN [24], RetinaNet [150], DETR [151], Mask-RCNN [23]	COCO [47], FLIR ADAS
		Traffic sign recognition		[91]	YOLO [17]	TT100K [44]
				[96]	Faster-RCNN [24], YOLOv3 [20], YOLOv5 [19]	ImageNet [46], KITTI [48], BDD100K [45]
				[90]	NanoDet4 [152]	TT100K [44]
				[92]	ResNet-50 [39], Microsoft Cognitive Services Image Captioning API [153]	ImageNet [46], TSRD [154]
				[93]	YOLOv3-tiny [20]	COCO [47]
			Camouflage	[54]	Faster-RCNN [24], YOLOv2 [18], SSD [21]	COCO [47], a STOP sign image
				[97]	Faster-RCNN [24], SSD [21], RetinaNet [150], CenterNet [155]	TT100K [44], a $40/h$ speed limit image
				[98]	3 proximity CNNs	LISA [43], GTSRB [42]
			Optical methods	[99]	YOLOv3 [20], Mask-RCNN [23], LISA-CNN, GTSRB-CNN, Google Vision API	LISA [43], GTSRB [42]
				[16]	VGG-16 [38], ResNet-50 [39]	ImageNet [46], real photos
				[101]	ResNet-50 [39]	ImageNet [46], real photos
				[102]	LISA-CNN, GTSRB-CNN	LISA [43], GTSRB [42]
				[100]	C3D [156], I3D [157], LRCN [158] , DN [159], TSN [160]	HMDB51 [161], UCF101 [162], UCF Crime [157]
				[103]	Inceptionv3 [41], VGG-19 [38], ResNet-50, ResNet-101 [39], GoogleNet [163], AlexNet [164], MobileNet [165], DenseNet [40]	ImageNet [46], real photos
				[15]	ResNet-50 [39]	ImageNet [46], real photos
		Trajectory prediction		[105]	SiamRPN [28], DaSiamRPN [29], DaSiamRPN+ [29]	VOT [166], DriveTruth [167], real video
			Trajectory	[107]	FQA [168], GRIP++ [30] , Trajectron++ [31]	Apolloscape [169], NGSIM, nuScenes [49]
				[106]	AgentFormer [170], Trajectron++ [31]	nuScenes [49]
		Vehicle detection		[11]	PointPillars [25], PointRCNN [26], PIXOR [27]	KITTI [48], real point cloud
			Adversarial objects	[109]	SSD_VGG [21], Faster-RCNN [24], YOLO [17], OpenALPR	Images from the internet
			Camouflage	[111]	Inceptionv3 [41], VGG-19 [38], ResNet-152 [39], DenseNet [40]	ImageNet [46], COCO [47], scene in CARLA
				[112]	Light-Head RCNN [171]	COCO [47], scene in CARLA
				[114]	SSD [21], Faster-RCNN [24], Mask-RCNN [23]	Scene in CARLA
				[113]	YOLOv5 [19], SSD [21], Faster-RCNN [24], Mask-RCNN [23]	COCO [47], scene in CARLA
			Patch/Sticker	[110]	YOLOv3 [20], YOLOv4 [172], YOLOv5 [19], Faster-RCNN [24]	Video from the internet
			Scene	[12]	PIXOR [27], VoxelNet [173], F-PointNet [174], PointPillars [25]	KITTI [48], real point cloud
				[115]	PointNet [174], SqueezeSeg [36], Cylinder3D [37], PointNet++ [35], PointASNL [175]	Semantic KITTI [142], real point cloud
			Sound waves	[117]	YOLOv3 [20], YOLOv4 [172], YOLOv5 [19], Faster-RCNN [24], Baidu Apollo	COCO [47], BDD100K [45], KITTI [48]
	Sensor	Camera perception	Optical methods	[122]	Mobileye 630 PRO, Tesla Model X	None
Victim vehicle	DNNs	LiDAR perception	Trajectory	[108]	PointPillar++ [176]	nuScenes [49]
		Traffic sign recognition	Patch/Sticker	[94]	Inceptionv3 [41], VGG-19 [38], ResNet-50, ResNet-101 [39], GoogleNet [163], AlexNet [164],MobileNet [165], DenseNet [40]	ImageNet [46], real photos
				[95]	YOLOv2 [18], Faster-RCNN [24]	LISA [43], MTSD [177], BDD100K [45]
			Optical methods	[104]	VGG-19 [38], ResNet-50 [39], GoogleNet [163], AlexNet [164],MobileNet [165], DenseNet [40]	ImageNet [46], real photos
	Sensor	Camera perception		[123]	SONY cameras, Tesla Model 3	None
		LiDAR perception		[121]	VLP-16 LiDAR, Baidu Apollo 5.0, PointPillars [25], Autoware	KITTI [48]
				[119]	VLP-16 LiDAR, RS-16 LiDAR, PointPillars [25], SECOND [178], Baidu Apollo	None
				[120]	VLP-16 PUCK LiDAR, PointPillars [25], PointRCNN [26], Baidu Apollo 5.0	KITTI [48]
		MSF perception		[124]	VLP-16 PUCK LiDAR, PointPillars [25], PointRCNN [26], PIXOR [27]	KITTI [48]
		Camera-based depth estimation		[125]	ZED camera, Intel RealSense camera, DJI drone, OpenCV(BM, SGBM) [179], DispNet [180], PSMNet [181], AANet [182]	None
		Traffic light recognition		[126]	AR0132AT evaluation board camera, Xiaomi dashcams, Hikvision camera, OpenMV H7 Devboard camera, Baidu Apollo	BDD100K [45]
		Traffic sign recognition		[67]	Aptina MT9M034 camera, Aptina MT9V034, Ring indoor security camera, LISA-CNN	LISA [43]
		Radar perception	mmWaves	[118]	Lincoln MKZ autonomous vehicle testbed (TI IWR6843 radar), Baidu Apollo	None