Benchmarking Physical Sensor Attack Robustness of Fusion-based Perception in Autonomous Driving

Zizhi Jin, Xiaoyu Ji, *Member, IEEE,* Yu Wang, Xuancun Lu, Yushi Cheng, Chen Yan, Wenyuan Xu, *Fellow, IEEE*

***Abstract*—As a safety-critical application, Autonomous Driving (AD) has received increasing attention from security researchers. AD heavily relies on sensors for perception. However, sensors themselves are susceptible to various threats since they are exposed to environments and vulnerable to malicious or interfering signals. To cope with situations where a sensor might malfunction, Multi Sensor Fusion (MSF) was proposed as a general strategy to enhance the robustness of perception models.**

In this paper, we focus on investigating MSF security under various sensor attacks and wish to answer the following research questions: (1) *Does fusion enhance security or not?* (2) *How does the architecture of the fusion model influence robustness?* To this end, we establish a rigorous benchmark for fusion-based 3D object detection robustness. Our new benchmark features 5 types of LiDAR attacks and 6 types of camera attacks. Different from traditional benchmarks, we take the physical sensor attacks into consideration during the corruption construction. Then, we systematically investigate 7 MSF-based and 5 single-modality 3D object detection models with different fusion architectures. Additionally, we provide insights and conduct feasibility experiments for enhancing the robustness of MSF-based models. We release the benchmarks and codes to facilitate future studies: https://github.com/Jinzizhisir/PSA-Fusion..

I. Introduction

In autonomous driving (AD), 3D object detection serves as the core basis of the perception stack, especially for the sake of path planning, motion prediction, collision avoidance, etc. LiDAR and camera are the two most important sensors for 3D object detection. LiDAR provides precise 3D spatial information through point cloud data, while cameras provide rich texture information through image data. The fusion of these two complementary sources of information is a common effort in both academia [1], [2], [3], [4], [5] and industry [6], [7], [8], [9], [10] to enhance perception performance.

However, many recent security studies [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] indicate that LiDAR and camera systems can be compromised by physical signals, e.g., laser, electromagnetic interference (EMI) and ultrasound. We adopt the term *physical sensor attacks* to describe attacks that employ physical signals to manipulate sensor output. The physical sensor attack will induce the point clouds and images inevitably to encounter significant corruption. As an extremely safety-critical application, autonomous driving particularly requires enhanced robustness to address the corruptions that may arise in the physical world.

Corresponding author: Xiaoyu Ji

Z. Jin, X. Ji, W. Yu, X. Lu, Y. Cheng, C.Yan and W.Xu are with the College of Electrical Engineering, Zhejiang University, Hangzhou, CN.

E-mail: zizhi, xji, 3200102036, xuancun_lu, yushicheng, yanchen, wyxu @zju.edu.cn

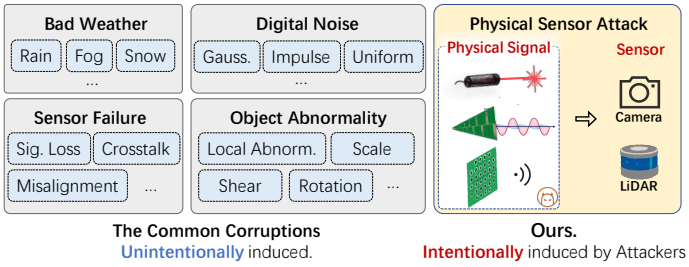


Fig. 1: **The common corruptions used in previous robustness benchmark and the sensor attack-based corruptions in this work.**

Many preceding works [23], [27], [22], [11] have considered sensor fusion as potential countermeasures. However, whether fusion can attenuate the attacks as anticipated remains an open question, lacking systematic research.

One common practice for such robustness analysis is to establish a benchmark [28]. Several benchmarks are proposed for image corruption [29], [30], [31] and point cloud corruption [32], [33]. As shown in Fig. 1, the corruptions used in those benchmarks can be grouped into bad weather, digital noise, sensor failure, object abnormalities, etc. Compare to those corruptions, the corruptions in this work are intentionally induced by attackers with physical sensor attack. Recently, a small amount of benchmarks [28], [34] has focused on the robustness of MSF-based perception. However, none of them has considered the corruption induced by sensor attack, and the number of MSF-based 3D object detection models under evaluation is still limited, e.g., only 3 models in [28] and 2 models in [34] are tested on the corrupted dataset.

In this paper, we propose a benchmark for evaluating the robustness of MSF-based 3D object detection against 11 types of sensor attacks. Such a benchmark could provide significant value to both academia and industry. As a shared reference, it could facilitate various activities, including developer training, assessing risks, and advancing the design of new MSF-based models. Based on the benchmark, we set out to answer the following research questions:

RQ1. Does fusion enhance security? Compared to single-modality models, can fusion models offer enhanced security? This is a fundamental and crucial question. The field of autonomous driving, being safety-critical, is highly susceptible to being targeted by attackers. However, no study has systematically investigated the robustness of multi-sensor fusion models in autonomous driving when faced with malicious sensor attacks. This study answers the question (detaied in Sec. V-B) by evaluating three aspects: targeted attack robustness, single source robustness, and overall robustness.

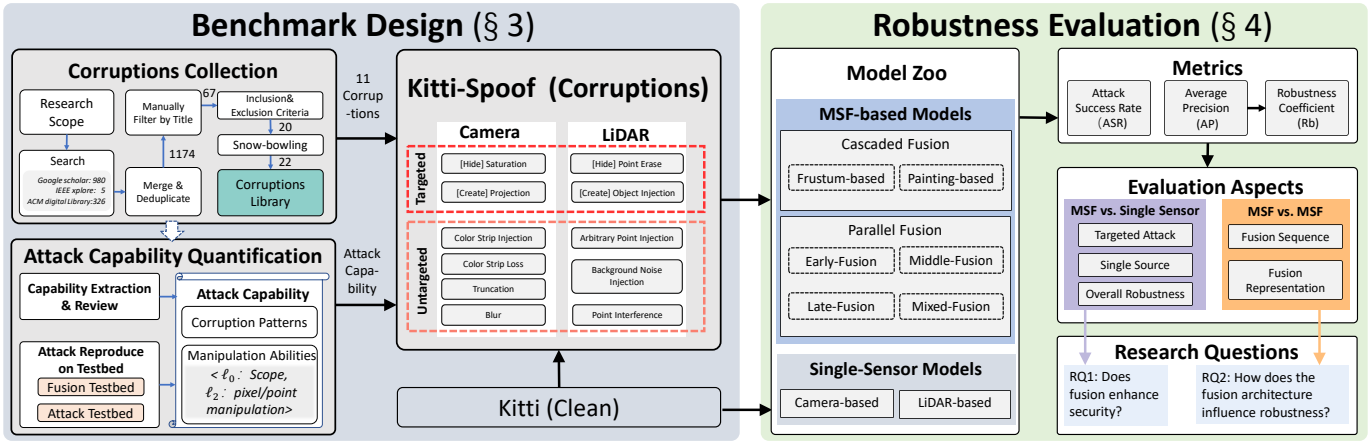


Fig. 2: **The benchmark overview.** First, we collect works related to sensor attacks as comprehensively as possible through a Systematic Literature Review (SLR) process. Second, we quantify the attack capability by reviewing the paper and reproducing the attacks on our physical testbed. Third, we design sensor attack corruptions for both LiDAR and camera sensors. By applying corruptions to typical autonomous driving datasets KITTI [35], we establish the sensor attack robustness benchmark dataset KITTI-spoof. Finally, We conduct large-scale experiments centering around two research questions to benchmark the robustness of MSF-based model against physical sensor attack.

RQ2. How does the architecture of the fusion model influence robustness? In the face of physical sensor attacks, do multi-sensor fusion-based detectors with varying architectures demonstrate performance disparities? If such differences exist, what are the underlying reasons? Previous research typically categorizes models into early, middle, and late fusion. However, we found that this classification method does not explicitly reveal the relationship between architecture and robustness. In this paper, we introduce a novel paradigm, categorizing models based on *fusion sequence* and *fusion representation*, and delve into the relationship between architecture and robustness using the concept of information entropy. Through this new classification approach, the answers to RQ2 have become clearer (detailed in Sec. V-C).

Building a benchmark to answer those questions is a challenging problem, especially when considering the physical feasibility and comprehensiveness of the dataset. Unlike purely digital corruptions, which allow arbitrary editing of images and point clouds, making the dataset physically realizable requires considering the capabilities of physical sensor attacks. However, previous sensor attacks have mainly focused on demonstrating their attack effectiveness but have not explicitly quantified their attack capabilities for benchmarking purposes.

To bridge this gap, we design the benchmark process as shown in Fig. 2. In summary, our contributions are concluded as follows:

- **Benchmark.** We present a large-scale robustness benchmark for MSF-based 3D object detection under physical sensor attack, namely Kitti-Spoof. The dataset contains 11 corruptions induced by laser, EMI, and acoustic.
- **Empirical Evaluation.** Based on the benchmark, we perform a large-scale (542,736 frames) empirical study to evaluate the sensor attack robustness on 7 MSF-based detectors and 5 single-modality detectors with different architectures.
- **Insights for Critical Research Question.** This paper systematically answers the fundamental and critical questions related to the robustness of MSF-based models.
- **Robustness Improvement.** We provide insights to enhance the robustness of MSF-based model. Based on the insights, we conduct feasibility experiments to improve the robustness of the SOTA MSF-based model.

II. Threat Model and Definition

A. Threat Model of Physical Sensor Attack

In this benchmark, we consider adversaries with the following assumptions.

Attack capability: The adversary conducts attacks outside the car to be stealthy. She can aim the camera or LiDAR and inject signals to attack them, and can solve the problem of aiming if needed

Sensor Assessment: The adversary has no direct access to the target sensors. She cannot physically touch them, alter the device settings, or install malware. However, we assume that she is fully aware of the characteristics of the target sensors. Such knowledge can be obtained from the user manual or by analyzing a sensor of the same model as the target sensor.

Black box: The adversary does not have access to the machine learning model or the perception system. Attackers can exploit only the characteristics and vulnerabilities of the sensors to achieve their attack target.

B. Scope and Definition of Sensor Attack Robustness

Firstly, we demarcate the scope of physical sensor attacks. Since sensors act as transducers that translate physical signals into electrical ones [36], we focus on physical signal attacks that corrupt the output of the sensor, with the threat model elaborated in Sec. II-A. The subsequent attacks do not fall within our benchmark's scope: (1) physical modification of the measured target, such as utilizing stickers [37], [38], [39], [40], [41], [42], [43] or 3D objects [44], [45], [46] to deceive sensors, and (2) attacking the digital transmission of sensor data in CAN bus [47], [48], or sensor networks [49], [50], [51].

We now define sensor attack robustness. To begin, we consider a detector $f : X \rightarrow Y$ trained on samples from

distribution $\mathcal{D}$. Most detectors are judged by their performance with the intersection of union (IoU) and a threshold (t) on test queries drawn from $\mathcal{D}$, i.e.,$\mathbb{P}_{(x,y)\sim\mathcal{D}}(IoU(f(x), y) > t)$. Yet in safety-critical applications, the detector may face malicious sensor attacks and is tasked with artificially corrupted inputs. In view of this, we suggest computing the detectors's *sensor attack robustness* $\mathbb{E}_{c\sim\mathcal{C}}[\mathbb{P}_{(x,y)\sim\mathcal{D}}(IoU(f(c), y) > t)]$, where $\mathcal{C}$ is a set of corruptions. The design of corruptions $\mathcal{C}$ should satisfy the physical realizability, i.e., $\|\mathcal{C} - X\| < \Phi$, where Φ is a set of physical attack capability of sensor attacks.

III. Benchmark Design

In this section, we first introduce the design methodology for the corrupted dataset KITTI-Spoof. We then detail the corruptions specific to the camera and LiDAR respectively.

A. Design Methodology for Kitti-Spoof

When designing the corrupted dataset Kitti-Spoof, we aim to ensure the comprehensiveness and physical feasibility of the dataset. The source papers and attack capability of corruptions are listed in Table I. Examples of corruptions are illustrated in Fig. 3.

1) Corruption Collection: We collected works related to sensor attacks with the scope defined in Sec. II-B as comprehensively as possible through a Systematic Literature Review (SLR) [52], [53] process. The SLR itself follows a three-step methodology comprising planning, conducting, and reporting, as depicted in Fig. 2. We define the search scope as "physically-realizable sensor attack" and use the following query to search for the terms in the documents.

> **Query**: ("physical" OR "real-world" OR "practical") AND "signal" AND ("attack" OR "vulnerability") AND ("LiDAR" OR "camera") AND ("autonomous driving" OR "self driving" OR "autonomous vehicle")

By leveraging the citation analysis software *Publish or Perish* [54], we collected studies from Google Scholar (980), IEEE Xplore (5), and ACM Digital Library (326). After removing duplicates from the total of 1311 search results, 1174 papers remained.

We only included studies related to physical sensor attacks on cameras or LiDARs. Moreover, we only included papers that first introduced the attack as well as those that made improvements to the attack. Initially, we conducted a preliminary filter based on paper titles, resulting in 67 potentially relevant papers. Subsequently, after reviewing the content, we shortlisted 20 articles. We then employed the snowballing technique on all these works to uncover resources overlooked in the initial search and applied the same inclusion criteria, leading to the addition of 2 new studies.

The *SLR* process yielded a total of 22 scientific works, and we distilled 11 types of corruptions from these papers.

2) Attack Capability Quantification: As shown in Fig. 2, we quantify the capabilities of sensor attacks in two steps. First, we extract useful information by reviewing source papers, which can ascertain the pattern characteristics. Some papers clearly describe the capability of attacks, while others do not. Second, we replicate each attack on our physical testbed, as depicted in Fig. 4, to ensure the physical feasibility of each attack and further clarify each attack's capability and limitations.

We quantify the capabilities of sensor attacks based on the corruption pattern characteristics and manipulation abilities. Similar to adversarial attacks [55], we utilize the ℓ_0 and l_2 norms to represent the manipulation abilities on images or point clouds, wherein l_0 norm signifies the attack scope, and l_2 norm illustrates the pixel/point manipulation capability. More specifically, for images, l_0 denotes the number (scope) of pixels that can be manipulated by the attack, and l_2 indicates how can the pixel values be manipulated. For point clouds, l_0 represents the number (scope) of points that can be affected by the attack, and l_2 signifies how can the distance of the points be manipulated.

Our physical testbed consists of a sensor fusion system and a signal transmission system. The fusion testbed, as shown in Fig. 5 in Appendix, comprises a Leopard USB3.0 camera [56] and a VLP-16 LiDAR [57] mounted on an Apollo-kit. The attack testbed, as shown in Fig. 4, includes a signal generator, an amplifier, and three types of signal transmitters, which can transmit laser, ultrasound, and electromagnetic signals.

Detailed attack capability quantification process of every corruption is described in Sec. III-C.

B. Corruptions Elaboration

1) Image Corruption: There are a total of six image corruptions in this benchmark, including two targeted corruptions and four untargeted corruptions.

[Camera-Hide]Laser-Saturation: The attack method [14], [15] involves using a high-power laser or a high-lumen light beam to directly irradiate the camera. This causes the light-sensitive module in the camera to be saturated, effectively hiding the real objects in the environment. The principle of this phenomenon is similar to that of overexposure in dynamic lighting conditions [58] in real life. Such overexposure is caused by excessive luminous flux and saturation of the light-sensitive module. The saturation (or overexposure) commonly occurs in real-world driving situations, such as when the car is coming out of a tunnel [59] or when an oncoming car activates its high-beam headlights [60]. In this situation, the camera's image will be overexposed and blinded.

[Create]Light-Projection: This attack method involves using a projector to cast images onto the environment [16], [17], [18], [19], [61] or directly projecting images into the camera [16]. While this attack method might seem somewhat naive, it represents a significant threat. Its effectiveness bears some similarity to sticker-based attacks. However, compared to such sticker attacks, it offers distinct harm. Firstly, it can be executed remotely without requiring the attacker to go over there and stick it himself. Secondly, it provides convenient control over the attack via signal manipulation. Thirdly, it can project elements into locations that are challenging for sticker-based attacks, such as trees by the roadside [17] or air [19]. Nevertheless, the primary drawback of this projection attack lies in its susceptibility to environmental lighting conditions.

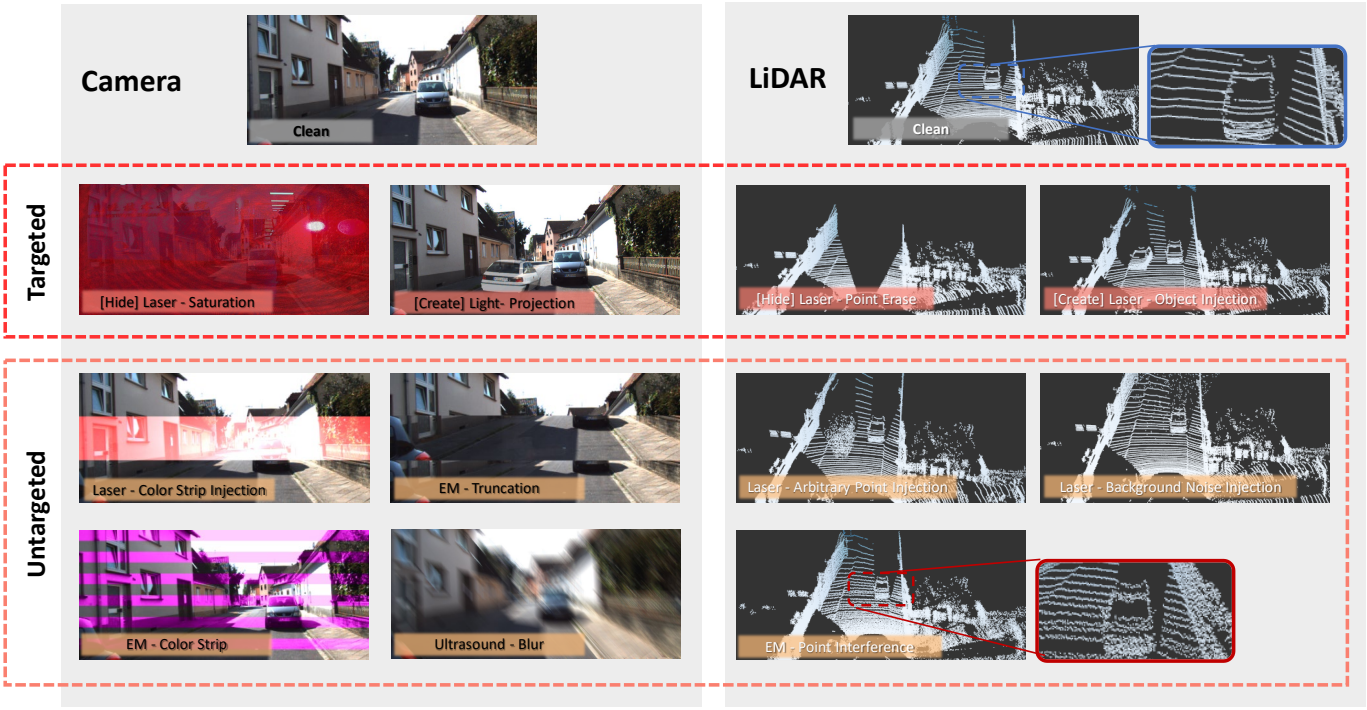


Fig. 3: **The corruptions in our benchmark.** The corruptions are named in the format *[attack target (if any)] signal-patterns*. There are 6 camera corruptions and 5 LiDAR corruptions. The corruptions are grouped into *targeted* and *untargeted* according to the attack effect on single-modality detectors. Best viewed on a screen and zoomed in.

Laser - Color Strip Injection: This attack method [20], [62] involves exploiting the rolling shutter of CMOS sensors, allowing attackers to inject a colored stripe. Prior research [20] evaluated the impact of this attack on traffic light recognition.

EM - Strip Loss and EM - Truncation: This attack [21] targets the camera interface bus used for image signal transmission and employs intentional electromagnetic interference (IEMI) to inject malicious signals, causing camera glitches. The principle of the attack is that cameras using MIPI CSI-2 transmission standard allocate a buffer for image signals. The start/end address of the buffer and the line pitch are passed to the Unicam (CSI Receiver). The image signals are transmitted by individual lines and decoded based on the fixed color filter arrangement. The camera will discard the lines that encounter transmission errors. If one line in the transmission is missing, it can disrupt the color interpretation of the subsequent lines during image processing, thereby causing color strips. If the start/end address of a buffer is missing, inter-frame content stitching appears, thereby causing truncation.

Ultrasound-Blur: This attack [22], [63] is based on a system-level vulnerability that image stabilizer hardware is susceptible to acoustic manipulation. By emitting deliberately designed acoustic signals, an adversary can control the output of an inertial sensor, which triggers unnecessary motion compensation and results in a blurred image.

2) Point Cloud Corruption: There are a total of five point cloud corruptions in this benchmark, including two targeted corruptions and three untargeted corruptions.

[Hide]Laser-Point Erase: Existing research has already demonstrated the feasibility of erasing point clouds using continuous-wave laser [25]and pulsed laser [23], [24], thereby hiding targeted objects. LiDAR functions by emitting lasers and receiving echoes from objects to perform time-of-flight measurements and distance measurements, ultimately generating point clouds. Existing point erasure methods fundamentally disrupt or hide the valid echoes from objects. Shin et al. [25] utilize a high-power (800mW) continuous laser to saturate the LiDAR's photodetectors, rendering them incapable of receiving valid echoes. Jin et al. [23] and Cao et al. [24] adopt pulsed lasers of specific frequencies to inject high-intensity points and then utilize the point cloud's echo filtering mechanism to filter out valid echoes.

[Create]Laser-Object Injection: This type of attack [23], [64], [65] employs a set of laser receiver and transmitter for controllable point cloud injection against mechanical LiDAR systems. The PLA-LiDAR [23] proved that it's feasible to inject point clouds in the physical world and directly spoof 3D object detection models using a black-box approach.

Laser - Arbitrary Point Injection: Several studies [14], [25], [27], [66], [67] have successfully implemented laser-based points injection attacks against LiDAR. However, these injected points exhibit a certain level of randomness rather than regular shapes shown in papers [23] and [64]. We suppose this might be due to differences in signal design and a lack of precise synchronization compared to controllable injection. Even though these attacks have not been proven to achieve targeted attack effects in the physical world, we are curious about their potential impact on fusion model performance.

Laser - Background Noise Injection: This type of attack [25] involves injecting random fake points using low-power lasers. The authors demonstrate that this may be due to the low-power laser causing an increase in baseline noise.

TABLE I: **The Attack Capability of The Transduction Attack Corruptions**

ID	Corruptions	Attack Capability			Source Paper
		Corruption Patterns	ℓ_0 :Scope	ℓ_2 :Pixel / Point Manipulation Quantification	
1	[Hide] Laser - Saturation	Global Exposure	All Piexels	Value addition on {R,G,B} channels according to quantum efficiency	[14], [15]
2	[Create] Light- Projection	Specified pattern	Specified location	Value superposition of projected pixels and original pixels	[16],[17],[18] [19],[61]
3	Laser - Color Strip Injection	color strip	Specific rows of the image.	Value addition on {R,G,B} channel according to quantum efficiency	[20], [62]
4	EM - Color Strip Loss	Uniform color strip	Specific rows of the image.	Filter array mismatch: G→R/B, R/B→G	[21]
5	EM - Truncation	Content Stitching	Specific rows of the image.	The image is stitched with the previous frame or next frame	[21]
6	Ultrasound - Blur	Linear Blur	All Piexels	Value superposition of a series of translated pixels	[22], [63]
7	[Hide] Laser - Point Erase	Point Erase	30° azimuth	The original points are erased	[23], [24]
8	[Create] Laser - Object Injection	Specified Object	20° azimuth	Fake points with random distance noise of 0.05 meters.	[23], [64], [65]
9	Laser - Arbitrary Point Injection	adversarial point cloud	30° azimuth	Fake adversarial points with raondom distance noise of 1 meters	[14], [25], [27] [66], [67]
10	Laser - Background Noise Injection	unfirom noise	30° azimuth	Fake random points within a distance of 100 meters	[25]
11	EM - Point Interference	sinusoidal noise	All Points	Sinusoidal noise wintin 0.05m added to the original points	[26]

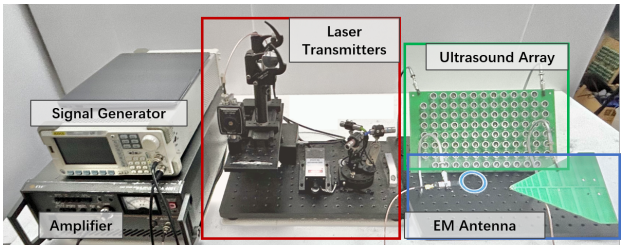


Fig. 4: **Attack Testbed.**

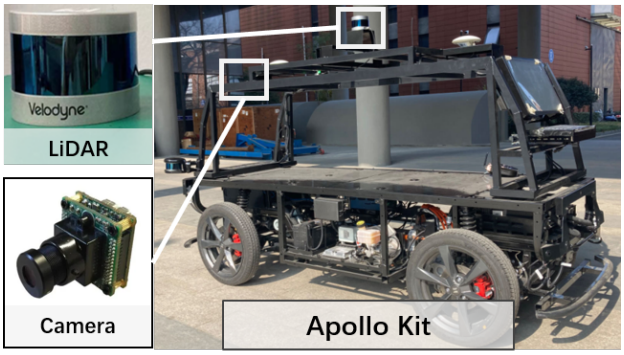


Fig. 5: **Fusion Testbed.**

In the everyday use of LiDAR, similar noise is sometimes observed, primarily due to interference between LiDARs, which is different from the principle of random noise injection attacks.

EM - Point Interference: This type of attack [26] exploits the susceptibility of time-of-flight (TOF) circuits to electromagnetic (EM) waves. By injecting EM signals at specific frequencies into the LiDAR's circuits, it disrupts the LiDAR's ranging function, consequently corrupting the global point cloud and introducing disturbances at radial distances.

C. Attack Capability Quantification

1) Image Corruption:

[Hide]Laser-Saturation: Previous papers have conducted the hiding attack on a camera using LED and laser, and the experimental results showed that laser can easily blind the camera even damage it compared to LED. In the previous experiments, the red laser (650 nm) was used. In order to make the experimental results more complete, in this paper, we use more wavelengths of lasers for the experiments, and we also tried high lumen beams. We discovered that at the same power, green lasers(550 nm) can cause more severe overexposure effects than lasers of other wavelength. This may be due to the higher proportion of green pixels in CMOS sensors. In fact, *[Hide]Laser-Saturation* represents the complete corruption or even erasure of image information. In this paper, we recorded the laser attack pattern against a white background and then added it to the {R,G,B} channels of the original image.

[Create]Light - Projection: We conduct tests with the two methods: projecting into the environment and directly projecting into the camera. However, achieving a successful direct projection into the camera proved challenging, as it necessitates precise optical focusing of the projector and high-precision aiming between the attack signal and the camera's photosensitive components. Through testing, we finally choose to implement the attack by projecting patterns into the environment. We find this approach convenient for launching *create* attacks and can effectively deceiving state-of-the-art 2D object detection models. However, the projection attacks can be notably challenging to execute successfully under strong lighting conditions.

Laser - Color Strip Injection: The authors of "Rolling Color" [20] extensively discuss the impact of pulsed lasers on images in their paper and provide a modeling method for Laser Attack. In this paper, we adopt their approach for designing

corruption.

EM - Strip Loss and EM - Truncation: The GlitcHike [21] demonstrates that attackers can utilize EMI to induce a color strip in images due to errors in the optical filters, such as the incorrect use of blue-green (B/G) and green-red (G/R) filters. As a result, the injected strip visually appears as a uniform shade of purple (distinguishing it from the uneven strip in *Color Strip Injection*). The Glitchhike paper provides evidence that attackers can control the position, width, and number of purple strips. Similarly, for truncation, attackers can adjust the signal to control the position of truncation. We have also confirmed this in testbed-based testing. Therefore, we follow the attack capability outlined in the paper for designing corruption.

Ultrasound - Blur: Based on the three types of pixel motions along different Degrees of Freedom (DOFs), the authors [22] categorize the blur patterns into linear blur, radial blur, and rotational blur. Through the physical experiments in our testbed, we have observed that linear blur is the most easily induced type of blur. Therefore, we adopt linear blur to design corruption.

2) LiDAR Corruption:

[Hide]Laser - Point Erase: In the paper [25] by Shin et al., they utilized an 800mW continuous laser to hide point clouds of a $41 * 42cm^2$ metal plate, but they did not quantify the specific attack capabilities. In our testbed experiments, we conducted experiments using 905nm lasers with power outputs of 200mW, 600mW, 1000mW, and 2000mW, respectively. We found that as the power increases, the effective range of removal also increases. Using a 2000mW laser, we were able to erase point clouds within approximately a $6° * 6°$ area. In the studies by Jin et al. [23] and Cao et al. [24], they conducted a detailed evaluation, demonstrating that attackers can remove target point clouds over a horizontal angle of more than 30° [23] and 40° [24], respectively. We have also confirmed this on our testbed. Considering the overall attack effectiveness and cost, we have decided to draw inspiration from the latter attack method for designing our corruption approach.

[Create]Laser - Object Injection: The authors of PLA-LiDAR [23] claim the capability to control up to 4000 points within 30°, a number sufficient for injecting point clouds of objects such as cars and pedestrians. In addition to the number points, to enhance the physical realizability, the position and shape control precision of points are crucial metrics need to be considered. Position precision refers to the attacker's ability to precisely control the overall position of the injected point clouds. Shape precision refers to the ability to maintain the injected points into specific shapes. Based on calculations of continuous data for 7 seconds (70 frames), we observed that the position and shape precision follows random uniform distribution. When we set the mean value is 0, the standard deviation of position precision is 48.2cm and standard deviation of the shape precision is 5.3cm. We take these two errors into consideration when designing the corruptions.

Laser - Arbitray Point Injection: In these previous works [14], [25], [27], [66], [67], the latest research claims the capability to inject up to 200 points. However, inspired by PLA-LiDAR, we believe that injecting thousands of points is feasible. Therefore, when designing corruption, we set the number of points to be the same as in *Create-Object Injection* and assign each point a mean error ranging from 0 to 1 meter to reflect its randomness.

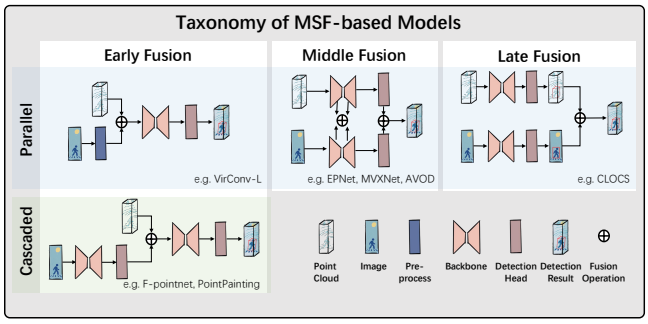


Fig. 6: **Three different fusion architectures for MSF-based 3D object detection models.**

Laser - Background Noise Injection: In the paper [25], Shin et al. claim the ability to inject noise within a 20° horizontal angle. However, this information alone is insufficient for designing corruption. Therefore, we conduct further experiments on our testbed using 5mW and 30mW continuous 905nm lasers at a distance of 7 meters. We find that it is feasible to inject noise within approximately a 30° horizontal angle range, and the noise is uniformly distributed within a range of 0 to 150 meters

EM - Point Interference: In the paper [26], Bhupathiraju et al. utilized EM signals at frequencies of 960.9MHz and 977.4MHz to induce sinusoidal and random patterns in the LiDAR's point cloud. Meanwhile, they achieved an average displacement of approximately 4cm under a 25dB EMI power. We follow the attack capability outlined in the paper for designing corruption.

IV. Models and Metrics

A. Fusion-based Model Collection

To collect as many appropriate SOTA MSF detectors as possible for our benchmark, we mainly focus on the MSF-related survey literatues [68], [1], [2], [3], [4], [5] and collect papers published in relevant top-tier conferences and journals during the last six years. Meanwhile, we refer to the 3D object detection leaderboard of Kitti benchmark [69] for the open-source models achieving SOTA performance. Eventually, we selected 7 state-of-the-art MSF systems as shown in Table. II.

The SOTA MSF detectors are mainly based on LiDAR-based 3D object detectors and try to incorporate image information into different stages of a LiDAR detection pipeline. According to different fusion stages, MSF models can be divided into *early fusion*, *middle fusion*, and *late fusion*(as shown in Fig. 6). There are three early-fusion models (F-Pointnet [70], Pointpainting [71] and VirConv-L [72]), two middle-fusion models (EPNet [73] and AVOD [74]), one late-fusion model (CLOCs [75]) and one mixed-fusion model (VirConv-T [72]) in our benchmark.

Most models process data from both sensors concurrently before the fusion operation. We define this concurrent approach as *parallel fusion*. While in early-fusion, a type of

TABLE II: **MSF-based 3D object detection model in our benchmark**

Model	Fusion Stage	Fusion Architecture	Fusion Representation		Fused Operator
			Camera Rep.	LiDAR Rep.	
F-Pointnet	Early	Cascaded	frustum	point cloud	region selection
PointPainting	Early	Cascaded	2D segmentation	point cloud	point-wise enhancement
VirConv-L	Early	Parallel	virtual points	point cloud	data concatenate
VirConv-T	Mixed	Parallel	virtual points	point cloud	data concatenate & ROI concatenation & box voting
EPnet	Middle	Parallel	image features	point features	point-wise attention
AVOD	Middle	Parallel	image features	BEV features	concatenation & MLP
Clocs	Late	Parallel	2D boxes	3D boxes	box consistency

fusion follows a sequential structure, which we refer to as *cascaded fusion*. In cascaded models, an image-based model is first used to obtain 2D recognition results, such as bounding boxes [70] or semantic classes [71]. These 2D results are then employed to enhance the point cloud, which is subsequently fed into the LiDAR-based model.

B. Metrics

We define robustness evaluation metrics in this section. 3D Object detection aims to locate, classify, and estimate oriented bounding boxes in the 3D space. The accuracy of object detection can be measured by IoU (Intersection over Union), which measures the intersection area between a ground-truth 3D bounding box B_g and a predicted 3D bounding box B_p over their union area.

We consider a successful detection for cars when IoU is larger than 0.7, which is the same as Kitti [35]. To better benchmark the robustness of models against different attacks, we have employed several advanced metrics based on IoU.

1) Attack Success Rate(ASR): ASR is employed to quantify the success rate of *targeted* attacks. In our benchmark, targeted attacks can achieve two types of effects against the black-box model: hide and create. The hide attack is considered successful solely when the target object evades detection. Conversely, the create attacks are deemed successful only when an initially non-existent object is generated within a designated region.

2) Average Precision (AP): AP approximates the shape of the Precision/Recall curve as:

$$AP\,|_{R_{40}} = \frac{1}{|R_{40}|} \sum_{r \in R_{40}} \max_{\tilde{r}:\tilde{r}>r} \rho(\tilde{r}) \tag{1}$$

where $\rho(r)$ gives the precision at recall r, meaning that instead of averaging over the actually observed precision values per point r, the maximum precision at recall value greater than or equal to r is taken. We adopted mean AP (mAP) [76], [77], by taking the average of APs at three difficulty levels (i.e., "Easy", "Moderate", and "Hard"), to measure the overall detection performance of a model.

3) Robustness(Rb): We define the Robustness of one MSF model on a corruption c as Rb_c:

$$Rb_c = \frac{mAP_c}{mAP_{clean}} \tag{2}$$

where mAP_c and mAP_{clean} represent the overall performance of the model on corruption *c* and clean data, respectively. The mean robustness of one model across multiple corruptions is denoted as mRb.

$$mRb = \frac{1}{|C|} \sum_{c \in C} Rb_c \tag{3}$$

V. Robustness Evaluation

A. Setup

We benchmark the 7 MSF-based models and 5 single-modality models using Kitti-Spoof. To ensure, as far as possible, that the models are compared on the same baseline, each model uses the official model parameters which are fine-tuned on the Kitti train dataset. Each model is tested on 12 datasets (1 clean + 11 corrupted), each dataset containing 3,769 LiDAR-camera frames. Since some of the models only support the detection of the "car" category, we calculate the performance of all models based on their detection results for cars. After obtaining the detection results for each frame, we compute the AP with a 0.7 IoU threshold. Then the AP results are used to calculate the robustness *Rb* (shown in Table IV). In addition, for targeted corruption, we calculate the attack success rates *ASRs* (shown in Table III). Based on these results, we engage in discussion and analysis centered around two research questions.

B. RQ1. Does fusion enhance security?

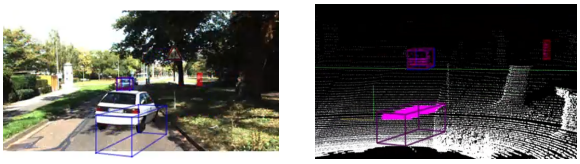
To comprehensively evaluate whether MSF-based models enhance security compared to single-modality models, we decompose security into the following three aspects:

1) ASR of Targeted Attack: Evaluating targeted attacks is of considerable importance due to its real-world relevance. In specific driving scenarios, attackers often aim to conceal target objects or create objects at predetermined locations, potentially inducing collisions or traffic jams as intended by the attacker. Past research has shown that single-modality models are notably susceptible to such targeted attacks, highlighting the unique and significant threats posed by targeted attacks that necessitate the attention of security researchers. Consequently, evaluating whether attackers can control the outputs of multi-modal models in a straightforward manner, similar to their control over single-modality models, serves as one of the key aspects in measuring model robustness.

We use the parameter ASR to evaluate targeted attack robustness. The results are shown in Table III. We take one camera-based detector ImVoxelNet [78] and one LiDAR-based

TABLE III: **Attack Success Rate**

Target Sensor	Attack Target	Corruption	Camera-only	LiDAR-only	Fusion Model						
			ImVoxelNet	PointPillar	F-pointnet	Pointpainting	virconv_l	virconv_t	Epnet	AVOD	CLOCS
Camera	Hide	Saturation	97.37%	\	92.58%	47.41%	0.04%	0.54%	7.78%	20.48%	88.51%
	Create	Projection	100.00%	\	96.77%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
LiDAR	Hide	Point Erase	\	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Create	Point Injection	\	100.00%	0.00%	95.93%	100.00%	98.27%	100.00%	100.00%	0.00%



(a) Image (b) Point Cloud.

Fig. 7: **The detection results of F-Pointnet against [Create]Light-Projection corruption. The ground point clouds are misdetected as a "car", indicating the success of the create attack.**

detector PointPillar [79] as the baseline. It can be observed that the 4 targeted attacks can achieve a high attack success rate against the single-modality detectors. In contrast, the ASR of the attacks on the fusion models varies significantly. Overall, **hiding is easier than creating**. Examining each type of attack, the LiDAR-Hide attack can successfully compromise all models, as the method of erasing point clouds can almost entirely eliminate an object's 3D information. This action consequently prevents the successful regression of the 3D bounding box. Continuing with this line of reasoning, Camera-create, which does not provide 3D information about the object, should logically be unable to succeed. This holds true for the majority of models. However, we were surprised to find that the Camera-Create attack successfully generated spoofed objects in F-pointnet. We provide an explanation by analyzing the model architecture of F-PointNet later. Moreover, the success rates of Camera-Hide attacks and LiDAR-Create attacks are inversely related. Models easily compromised by Camera-Hide attacks are less susceptible to LiDAR-Create attacks and vice versa. This indicates that existing fusion models often rely more on one sensor source (majority). The late fusion in CLOCs treats detection results from both modalities equally, eliminating this bias. However, due to the structural characteristics of CLOCs, it tends to prune rather than create new discoveries, which can be hazardous in real autonomous driving scenarios.

We provide further explanations for the attack results, analyzing them on a model-by-model basis.

F-Pointnet. Due to the cascaded fusion structure of F-Pointnet, if an object is not detected in the image, then the object's point cloud will be filtered out, leading to a high success rate for the Camera-Hide attack. It is important to note that, strictly speaking, the Camera-Create attack cannot succeed in all models since merely modifying an image does not provide 3D information. However, we were surprised to find that the Camera-Create attack successfully generated spoofed objects in F-Pointnet. To understand why the Camera-Create attack succeeds, we visualized the detection results as shown in Fig. 10. We found that the created objects in F-Pointnet are instances where the ground is mis-detected as a car. We suppose that after the filtering mechanism in F-Pointnet, the ground point cloud obtained features closely resembling those of a car roof.

Pointpainting. Camera-Hide and LiDAR-Create can achieve 47.41% and 95.93% ASRs, respectively, on Pointpainting. The PointPainting architecture consists of three main stages. (1) an image based semantic segmentation network which computes the pixel-wise segmentation scores; (2) a fusion stage that LiDAR points are painted with semantic segmentation scores; and (3) a LiDAR-based 3D detection network. According to the taxonomy in this paper, it is an early fusion. The image segmentation scores are appended to the LiDAR point to create the painted points. The paited points can be consumed by any LiDAR network that learns an encoder, since PointPainting just changes the input dimension of the LiDAR points. In this benchmark, we used the PoinPillar and decorate the point cloud with the semantic segmentation scores for 4 classes in Kitti. Therefore, both the image's classification information and the raw point cloud information have the potential to influence the final detection results. Furthermore, based on the results, it can be observed that the fusion architecture of point painting exhibits greater robustness against camera attacks compared to LiDAR attacks.

EPNet. EPNet is very robust against Camera-Hide (ASR 7.78%) but is vulnerable to LiDAR-Create(ASR 100%). EPNet consists of a two-stream backbone network, composed of a geometric stream and an image stream. The two streams produce the point features and semantic image features, respectively. In image stream, EPNet adopts four cascaded $3*3$ convolutional blocks to extract image semantic features in different scales, denoted as $F_i(i = 1, 2, 3, 4)$. The geometric stream comprises four paired set abstraction $S_i(i = 1, 2, 3, 4)$ and feature propagation layers $P_i(i = 1, 2, 3, 4)$ [80] for feature extraction. The point features S_i are combined with the image features F_i with the aid of LI-Fusion module. Overall, the Fusion module allows point features and image features to be deeply fused in the backbone network. However, as the primary role of the image is to enhance the point cloud, point cloud features still play a predominant role throughout the pipeline. Thus, it is easier for LiDAR-based corruptions to compromise EPNet compared to camera-based corruptions.

AVOD. AVOD is robust against Camera-Hide (ASR 20.48%) but is vulnerable to LiDAR-Create(ASR 100%). In AVOD, the image and bird-eye-view (BEV) point feature maps, which are generated by feature extractors, are fused in region proposal network(RPN) Both feature maps are then used by the RPN to generate non-oriented region proposals, which are passed to the detection network for dimension refinement, orientation estimation, and category classification.

TABLE IV: **The Robustness(Rb) of 5 single-modality Detectors and 7 MSF-based on Kitti-Spoof.**

Target Sensor	Corruption	Camera-only		LiDAR-only			Fusion Model						
		ImVoxelNet	SMOKE	Second	PointPillar	3DSSD	F-PointNet	PointPainting	VirConv_L	VirConv_T	EPNet	AVOD	CLOCs
Camera	[Hide] Laser - Saturation	0.415	0.069	/	/	/	0.226	0.402	**0.999**	0.995	0.804	0.592	0.315
	[Create] Light- Projection	0.668	0.852	/	/	/	0.467	0.973	0.999	**1.000**	0.999	0.995	0.984
	Laser - Color Strip Injection	0.520	0.203	/	/	/	0.962	0.832	**0.999**	0.993	0.797	0.752	0.993
	EM - Color Strip	0.549	0.749	/	/	/	0.947	0.916	0.967	**0.985**	0.891	0.790	0.992
	EM - Truncation	0.010	0.000	/	/	/	0.080	0.330	**0.999**	0.933	0.782	0.404	0.320
	Ultrasound - Blur	0.001	0.000	/	/	/	0.386	0.330	**0.967**	0.958	0.790	0.411	0.636
LiDAR	[Hide] Laser - Point Erase	/	/	0.655	0.645	0.661	0.597	0.638	0.653	0.676	**0.683**	0.611	0.665
	[Create] Laser - Object Injection	/	/	0.781	0.778	0.767	0.775	**0.890**	0.793	0.796	0.707	0.830	0.889
	Laser - Arbitrary Point Injection	/	/	0.893	0.873	0.894	0.784	0.892	0.890	**0.910**	0.884	0.875	0.888
	Laser - Background Noise Injection	/	/	0.814	0.855	0.742	0.516	0.898	0.854	0.922	0.729	0.839	**0.959**
	EM - Point Interference	/	/	0.979	0.987	0.981	0.960	0.985	0.971	**0.994**	0.993	1.001	0.993
Mean Robustness on Camera Cor. (mRb^C)		0.359	0.312	/	/	/	0.511	0.630	**0.988**	0.977	0.844	0.657	0.707
Mean Robustness on LiDAR Cor. (mRb^L)		/	/	0.825	0.827	0.809	0.726	0.861	0.824	0.850	0.799	0.831	**0.879**
Mean Robustness on All Corruptions (mRb)		0.650	0.625	0.920	0.922	0.913	0.609	0.735	0.918	**0.923**	0.824	0.737	0.785

We can observe that the fusion architecture of AVOD puts the features of images and point clouds on an equal footing, which distinguishes it from EPNet. In EPNet, image features are used to aid in enhancing point cloud features. Although the AVOD architecture is intended to be equal for camera and LiDAR features, it is apparent that after training, the model is biased towards relying on the LiDAR features. This is evidenced by the attack results, where mean ASRs of camera attack vs LiDAR attack are 10.24% vs 100%.

CLOCs. Camera-Hide (ASR 88.51%) is able to successfully compromise CLOCs, whereas LiDAR-Create (ASR 0%) isn't. CLOCs is a late fusion approach that merges camera and LiDAR detection candidates before applying Non-Maximum Suppression (NMS). CLOCs employs significantly reduced thresholds for each sensor to optimize their recall rate. If the 2D and 3D bounding boxes have a large enough IoU in the image plane, then their information will be combined into a single tensor for subsequent processing. However, 3D (or 2D) bounding boxes for which no matching 2D (or 3D) bounding box can be found will be ignored. Overall, CLOCs, like most late fusions, tend to prune rather than create new discoveries, which explains why hide attacks are easy to succeed while create attacks are difficult.

Observation 1 (RQ1): For camera attacks, all fusion models (7/7) can reduce the attack success rate. However, no fusion model (0/7) can defend against the LiDAR-Hide attack, and only some models (4/7) can attenuate the LiDAR-Create attacks. The black-box targeted attacks, originally designed against single-modality detectors, still retain the potential capability to compromise fusion models. However, selecting the right fusion model can effectively mitigate the impact of such attacks.

2) RQ1.2 Single Source Robustness: Single Source Robustness refers to the average robustness of a model when facing attacks on a single sensor (such as a camera or LiDAR). Since single-modality models are only exposed to attacks targeting the sensor they use, single-source robustness allows for a comparison of the robustness between MSF and single-modality models when confronting the same attacks.

We utilize the parameters mRb^C and mRb^L to represent the model's mean robustness under camera or LiDAR corruption, respectively. The results are shown in Table IV, We found that the mRb^C of all MSF models surpasses that of camera-based models. This improvement can be attributed in part to the fusion of point clouds, which effectively enhances robustness. Another contributing factor is the generally inferior performance of existing open-source camera-based models, leading to their lower robustness. In contrast, only 4 out of the 7 MSF models in our experiments showed a superior mRb^L to LiDAR-based models. This suggests that fusion doesn't necessarily guarantee enhanced robustness, and selecting the right fusion method requires additional effort.

Observation 2 (RQ1): When considering single source attacks: Compared to the camera-based model, all fusion models (7/7) can enhance the robustness against camera attacks. Compared to LiDAR-based models, most fusion models (5/7) can increase robustness against LiDAR attacks.

3) RQ1.3 Overall Robustness: Overall robustness refers to the model's mean robustness under all corruptions in this benchmark. Given the increased diversity of sensors in the MSF-based models, they are exposed to a greater number of potential attack vectors. This aspect is pivotal and cannot be ignored when evaluating robustness and security.

We use the parameter mRb to evaluate overall robustness. For Camera-based models, we set their robustness to all LiDAR corruptions as 1, and vice versa for the LiDAR-based models. From Table IV, we observe that the majority of MSF models have greater mRb compared to camera-based models. This discrepancy may be attributed to the subpar performance of existing open-source camera-based models. The best mRb is exhibited by the LiDAR-based model. Meanwhile, the SOTA MSF models, VirConv-L and VirConv-T, also demonstrate commendable mRb. Overall, MSF-based models have a lower mRb compared to LiDAR-based models.

Observation 3 (RQ1): Since MSF-based models are exposed to sensor attacks from both sensors, the majority (6/7) of existing fusion models do not enhance overall robustness.

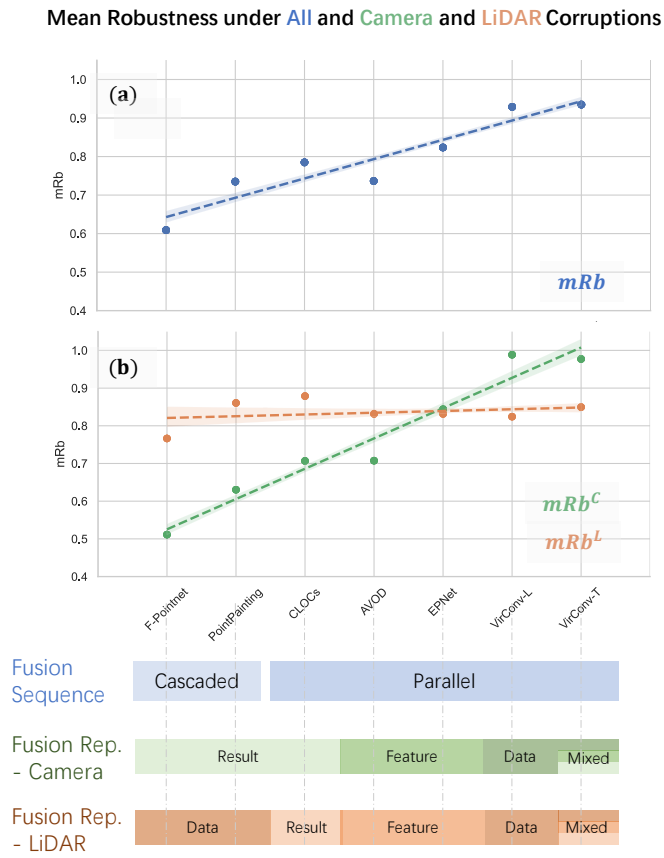


Fig. 8: **The mean robustness under (a) all corruptions and (b) camera and LiDAR corruptions.** The models on the X-axis are ranked based on two criteria. The first criterion is the fusion sequence, in the order of [cascaded, parallel]. Within the same fusion sequence, the information entropy of fusion representation serves as the second criterion, in the order of [result, feature, data].

Answer to RQ1

Considering *targeted attack robustness*, *single source robustness*, and *overall robustness*, most MSF-based models demonstrate enhanced robustness compared to camerabased models but exhibit weaker robustness when compared to LiDAR-based models. However, the state-of-the-art fusion model, such as VirConv-T, is expected to enhance robustness across all aspects, showcasing the potential of MSF in improving robustness.

C. RQ2. How does the architecture of the model influence robustness?

To answer RQ2, we compare MSF-based models with each other, aiming to evaluate which fusion design is more robust. We approach this comparison from two perspectives: fusion sequence and fusion data representation.

1) Fusion Sequence: According to the results in Table IV, the robustness of a model varies greatly with different architectures. To discern the relationship between fusion sequence and robustness, we categorize the models into cascaded fusion and parallel fusion. As the mRb under all corruptions shown in Fig. 8, we observed that the robustness of cascaded fusion is generally lower than that of parallel fusion. We suppose this is due to the cascade effect, where errors caused by corruption in a single sensor propagate throughout the detection pipeline, subsequently reducing overall robustness. In contrast, parallel fusion allows data from the two sensors to reinforce each other, thereby enhancing robustness.

Observation 4(RQ2) From the perspective of the fusion sequence, parallel fusion exhibits better robustness than cascaded fusion.

2) Fusion Representation: From the input to the output of the detection pipeline, data representation transitions from original *data* to *feature* and then to *results*. We adopt information entropy, denoted as $\mathcal{H}_X$, to intuitively quantify the information content of data X. In neural networks, basic operations such as convolution, activation functions, ROI pooling, NMS, and FC can lead to information loss [81]. Thus, we can intuitively derive the following relationship:

$$\mathcal{H}_{data} > \mathcal{H}_{feature} > \mathcal{H}_{result} \tag{4}$$

Additionally, We use $\mathcal{H}_{FR}(M)$ to represent the information entropy contained in the Fused Representation of model M.

First, let's consider the cascaded models: F-Pointnet and PointPainting. These two models both use 2D results generated from the image to fuse with the original point cloud. While F-Pointnet reduces the information entropy of the point cloud by filtering the point cloud using 2D detection results. we have:

$$\mathcal{H}_{FR}(Pointpainting) > \mathcal{H}_{FR}(F-pointnet) \tag{5}$$

Second, let's consider the parallel fusion models VirConv-T, VirConv-L, EPNet, AVOD, and CLOCs. The fused representations of the five models are shown in Fig. 8. It's important to note that the LiDAR input of AVOD is a bird's-eye view (BEV). Clearly, the information entropy of BEV is less than that of the original point cloud. Thus, we can determine that $\mathcal{H}_{FR}(AVOD)$ is less informative than $\mathcal{H}_{FR}(EPNet)$, but we cannot compare $\mathcal{H}_{FR}(AVOD)$ and $\mathcal{H}_{FR}(CLOCs)$. Therefore, we have:

$$\begin{aligned}\mathcal{H}_{FR}(VirConv-T) > \mathcal{H}_{FR}(VirConv-L) > \\ \mathcal{H}_{FR}(EPNet) > \mathcal{H}_{FR}(AVOD), \mathcal{H}_{FR}(CLOCs)\end{aligned} \tag{6}$$

As shown in Fig. 8(a), the overall robustness of the models also follows the information entropy relationship in Equ. 5 and Equ. 6. This confirms the *Observation 5*.

Observation 5 (RQ2): In general, given the same fusion sequence, the more comprehensive the information contained in the fused representation, the stronger the robustness. The comprehensiveness of information is ranked as data > feature > results.

Further analysis, as shown in Fig. 8(b), reveals that different fusion representations primarily influence the mRb^C. Moreover, the greater the information entropy ($\mathcal{H}_{FR}$), the stronger the mRb^C. However, the variation of $\mathcal{H}_{FR}$ has very few impacts on the mRb^L. This is because those MSF models are based on point cloud-based 3D object detectors and incorporate image information into various stages of the detection pipeline.

Answer to RQ2

Overall, different fusion sequences and fusion representations influence robustness with the following characteristics:
1) Parallel fusion exhibits better robustness than cascaded fusion. 2) The more comprehensive the information contained in the fused representation, the greater the robustness. The comprehensiveness of information is ranked as data > feature > results. 3) Different fusion architectures primarily influence the robustness to camera corruption.

VI. Robustness Improvement

A. Insights for Robustness Improvement

Based on the answers of RQ1 and RQ2, we provide insights for enhancing robustness. A Fusion architecture exhibiting the following characteristics can enhance robustness against physical sensor attacks: 1) Independence, 2) Parallel Fusion, and 3) Data Fusion.

Independence refers to the ability of each modality to achieve the final 3D object detection independently of the others. This allows one modality to potentially complete the final task even when subjected to data erasure attacks, such as the camera-hide and lidar-create discussed in this paper.

Parallel Fusion refers to incorporating sensor data equitably into the detection model during fusion, rather than designating one sensor as primary and another as auxiliary. Empirical evidence has shown that if bias is introduced towards one modality during fusion, the model becomes more vulnerable to attacks targeting the primary sensor. Moreover, the auxiliary sensor, having insufficient weight, faces challenges in effectively correcting the outcomes.

Data Fusion suggests that integration should occur at the raw data level, rather than at the feature or result levels. This is because raw data retains more comprehensive information, and experimental findings have affirmed the increased robustness of data fusion compared to other fusion methods. However, when fusing LiDAR and camera data, the process encounters challenges due to the heterogeneous nature of point clouds and images, which impedes direct data fusion.

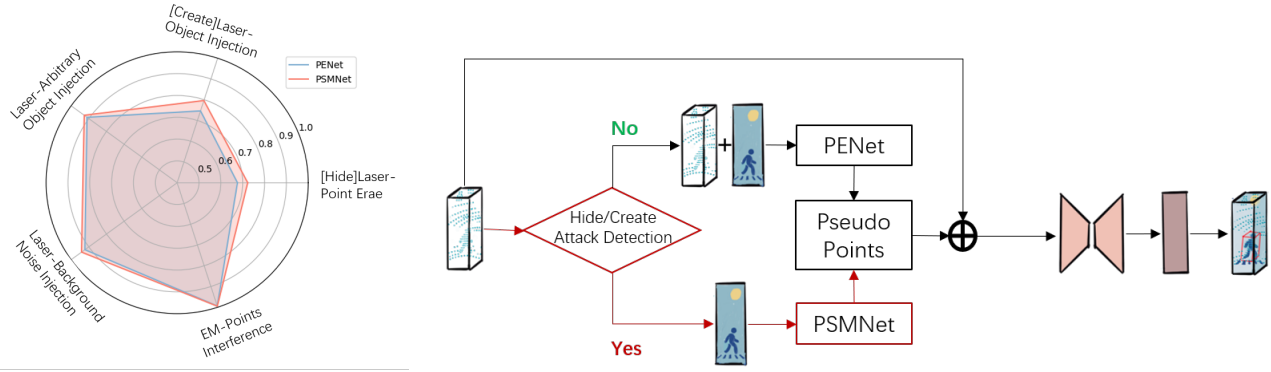
B. Feasibility Experiment for Robustness Improvement

Based on the evaluation experiments in Sec. V, we find that the SOTA model Virconv-T exhibits the highest level of robustness among the seven MSF-based models considered. However, the results presented in Table IV indicate that Virconv-T's robustness is not entirely satisfactory: it is still vulnerable to *[Hide]Laser - Point Erase* and *[Create]Laser - Object Injection* (abbreviated as *hide* and *create* in this section). Therefore, we are motivated to leverage our insights to enhance the robustness of Virconv-T. Clearly, improving the robustness of an already highly robust model presents a challenge.

Firstly, we analyze the model structure to explain why the LiDAR *hide* and *create* attacks significantly affect the performance of VirConv. VirConv employs a data fusion approach where the image is converted into pseudo points through PENet, and then these pseudo points are fused with real point cloud data. However, the two data modalities in Virconv are not independent. Since PENet generates pseudo points from image with the guidance of the real point cloud, making the point cloud capable of influencing both modalities simultaneously. As shown in Fig. 9, when facing LiDAR hide or create attacks, the pseudo points generated by PENet will be severely corrupted. Therefore, in Virconv-T, the point cloud has become the dominant data, and the image alone lacks the capability to independently perform the 3D object detection task. While this approach has the advantage of defending against image attacks using point cloud data, it potentially amplifies the effects of LiDAR attacks. This motivates us to enhance VirConv by introducing independence between the data from two sensors and incorporating mechanisms that enable VirConv to flexibly utilize data from different modalities, thereby improving its robustness.

Fig. 9: **Pseudo Points generated by PENet and PSMNet.** (Best Viewed on a screen and zoomed in) When facing LiDAR attacks, the pseudo points generated by PENet will be corrupted. However, the pseudo points generated by PSMNet are not affected by the attack and retain high similarity to the benign point cloud.

To improve the robustness of VirConv-T, we utilize PSMNet [82], which can generate virtual points solely based on image information, endowing the image channel with the capability to independently perform 3D object detection and correct the final detection results. As shown in Fig. 9, PSMNet maintains similarity to the benign point cloud even under LiDAR attack. We compare the pseudo points generated by PSMNet and PENet on VirConv-T. As shown in Fig. 10(a), the results demonstrate that PSMNet enhances VirConv-T's robustness against LiDAR corruptions, especially LiDAR *hide* and *create* attacks. Therefore, to enhance robustness while preserving the original advantages of VirConv-T, as shown in Fig. 10(b), we designed an adaptive attack detection mechanism. The attack detection method is based on the LIFE [83] approach to determine whether it is facing LiDAR *hide* and



(a) PSMNet can improve robustness in terms of LiDAR hide and create attacks.

(b) The architecture of VirConv-Rb.

Fig. 10: **VirConv-Rb improves robustness by attack detection and PSMNet**

TABLE V: **The mean robustness (mRb) of 7 previous MSF-based models and VirConv-Rb on Kitti-Spoof.**

Model	F-ConvNet	Point Painting	CLOCs	AVOD	EPNet	VirConv-L	VirConv-T	**VirConv-Rb (Ours)**
mRb	0.609	0.735	0.785	0.737	0.824	0.918	0.923	**0.934**

create attack. The detection method can achieve a success rate of 100% and 95.2% for detecting *hide* and *create* attacks, respectively. After attack detection, we incorporate a new channel utilizing PSMNet into our model to generate pseudo points upon detection of a LiDAR attack. We named this new model, which incorporates attack detection and modality independence, VirConv-Rb.

We evaluate the performance of VirConv-Rb on 11 corruptions and calculate the mean robustness. As shown in Table. V, VirConv-Rb achieves the highest robustness compared to other seven MSF-based models in this paper. This mainly due to VirConv-Rb improves the robustness against *[Hide]Laser - Point Erase* and *[Create]Laser - Object Injection*. Meanwhile, VirConv-Rb preseves the high robustness of VirConv against camera-based corruptions.

VII. Related Work

Understanding and analyzing the robustness of fusion-based perception has been broadly studied with digital data corruption (e.g. occlusion [84], [85], noise [86], [85] or downsampling [86]) and worst-case adversarial perturbation [87], [86], [88]. While these works bring intriguing results in most cases, they share two limitations. First, the corrupted data they adopted is purely digital, not reflecting the challenges fusion systems might encounter in the real world. Second, the number of fusion models they tested is limited (fewer than 3) and the performance of models is not state-of-the-art. This could potentially undermine the validity of the empirical conclusions drawn, which may lead to contradictory conclusions in different works. For instance, [84] said *"the later the sensor data is fused, the greater the detection rate of the object detectors is"*, while [88] said *"early fusion is more robust than late fusion"*.

VIII. Conclusion

In this paper, we introduce, to our knowledge, the first comprehensive benchmarks for MSF-based Models under physical sensor attacks by introducing a new dataset including 11 types of sensor attacks. We designed and conducted a rigorous SLR and attack capability quantification to ensure the comprehensiveness and physical feasibility of Kitti-Spoof as much as possible. Based on evaluating 542,736 frames on 7 MSF-based models and 5 single-modality models, we answer two open research questions: *RQ1) Does fusion enhance robustness?* We find most fusion models reduced overall robustness when considering attacks from both sensors. This challenges the consistent understanding of previous research. *RQ2) How does the architecture of the model influence robustness?* We adopted a novel paradigm to categorize models and introduced the concept of information entropy, which surprisingly revealed the relationship between model architecture and robustness. Finally, we provided some insights for enhancing robustness. We hope that our benchmark can aid in improving the performance of MSF-based models. The study can serve as a reference for researchers concerned with the security of MSF-based models in safety-critical applications.

References

[1] J. Fayyad, M. A. Jaradat, D. Gruyer, and H. Najjaran, "Deep learning sensor fusion for autonomous vehicle perception and localization: A review," *Sensors*, vol. 20, no. 15, p. 4220, 2020.
[2] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.
[3] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, and D. Cao, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, 2021.
[4] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: a review and new outlooks," *arXiv preprint arXiv:2206.09474*, 2022.
[5] R. Qian, X. Lai, and X. Li, "3d object detection for autonomous driving: a survey," *Pattern Recognition*, vol. 130, p. 108796, 2022.
[6] "Waymo driver," https://waymo.com/waymo-driver/, 2023.
[7] "Cruise," https://getcruise.com/, 2023.
[8] "Mobileye," https://www.mobileye.com/solutions/, 2023.
[9] "Aptiv," https://www.aptiv.com/, 2023.
[10] "Apollo," https://github.com/ApolloAuto/apollo, 2023.

[11] C. Yan, H. Shin, C. Bolton, W. Xu, Y. Kim, and K. Fu, "Sok: A minimalist approach to formalizing analog sensor security," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 233–248.

[12] Y. Xu, X. Han, G. Deng, J. Li, Y. Liu, and T. Zhang, "Sok: Rethinking sensor spoofing attacks against robotic vehicles from a systematic view," in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2023, pp. 1082–1100.

[13] J. Shen, N. Wang, Z. Wan, Y. Luo, T. Sato, Z. Hu, X. Zhang, S. Guo, Z. Zhong, K. Li *et al.*, "Sok: On the semantic ai security in autonomous driving," *arXiv preprint arXiv:2203.05314*, 2022.

[14] J. Petit, B. Stottelaar, M. Feiri, and F. Kargl, "Remote attacks on automated vehicles sensors: Experiments on camera and lidar," *Black Hat Europe*, vol. 11, no. 2015, p. 995, 2015.

[15] C. Yan, W. Xu, and J. Liu, "Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle," *Def Con*, vol. 24, no. 8, p. 109, 2016.

[16] 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*, 2020.

[17] B. Nassi, D. Nassi, R. Ben-Netanel, Y. Mirsky, O. Drokin, and Y. Elovici, "Phantom of the adas: Phantom attacks on driver-assistance systems." *IACR Cryptol. ePrint Arch.*, vol. 2020, p. 85, 2020.

[18] G. Lovisotto, H. Turner, I. Sluganovic, M. Strohmeier, and I. Martinovic, "Slap: Improving physical adversarial examples with short-lived adversarial perturbations." USENIX, 2021.

[19] 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.

[20] C. Yan, Z. Xu, Z. Yin, X. Ji, and W. Xu, "Rolling colors: Adversarial laser exploits against traffic light recognition," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1957–1974.

[21] Q. Jiang, X. Ji, C. Yan, Z. Xie, H. Lou, and W. Xu, "Glitchhiker: Uncovering vulnerabilities of image signal transmission with iemi."

[22] X. Ji, Y. Cheng, Y. Zhang, K. Wang, C. Yan, W. Xu, and K. Fu, "Poltergeist: Acoustic adversarial machine learning against cameras and computervision," *algorithms*, vol. 47, no. 26, p. 11.

[23] 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 (SP)*. IEEE Computer Society, 2022, pp. 710–727.

[24] 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 (USENIX Security 23)*, 2023, pp. 2993–3010.

[25] H. Shin, D. Kim, Y. Kwon, and Y. Kim, "Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications," in *International Conference on Cryptographic Hardware and Embedded Systems*. Springer, 2017, pp. 445–467.

[26] S. H. V. Bhupathiraju, J. Sheldon, L. A. Bauer, V. Bindschaedler, T. Sugawara, and S. Rampazzi, "Emi-lidar: Uncovering vulnerabilities of lidar sensors in autonomous driving setting using electromagnetic interference," in *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2023, pp. 329–340.

[27] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, "Adversarial sensor attack on lidar-based perception in autonomous driving," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 2267–2281.

[28] X. Gao, Z. Wang, Y. Feng, L. Ma, Z. Chen, and B. Xu, "Benchmarking robustness of ai-enabled multi-sensor fusion systems: Challenges and opportunities," *arXiv preprint arXiv:2306.03454*, 2023.

[29] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[30] C. Kamann and C. Rother, "Benchmarking the robustness of semantic segmentation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8828–8838.

[31] S. F. Altindis, Y. Dalva, H. Pehlivan, and A. Dundar, "Benchmarking the robustness of instance segmentation models," *arXiv preprint arXiv:2109.01123*, 2021.

[32] S. Li, Z. Wang, F. Juefei-Xu, Q. Guo, X. Li, and L. Ma, "Common corruption robustness of point cloud detectors: Benchmark and enhancement," *arXiv preprint arXiv:2210.05896*, 2022.

[33] X. Yan, C. Zheng, Z. Li, S. Cui, and D. Dai, "Benchmarking the robustness of lidar semantic segmentation models," *arXiv preprint arXiv:2301.00970*, 2023.

[34] Y. Dong, C. Kang, J. Zhang, Z. Zhu, Y. Wang, X. Yang, H. Su, X. Wei, and J. Zhu, "Benchmarking robustness of 3d object detection to common corruptions in autonomous driving," *arXiv preprint arXiv:2303.11040*, 2023.

[35] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[36] K. Fu and W. Xu, "Risks of trusting the physics of sensors," *Communications of the ACM*, vol. 61, no. 2, pp. 20–23, 2018.

[37] 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.

[38] Y. Zhao, H. Zhu, R. Liang, Q. Shen, S. Zhang, and K. Chen, "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 1989–2004.

[39] Z. Kong, J. Guo, A. Li, and C. Liu, "Physgan: Generating physical-world-resilient adversarial examples for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 254–14 263.

[40] L. Huang, C. Gao, Y. Zhou, C. Xie, A. L. Yuille, C. Zou, and N. Liu, "Universal physical camouflage attacks on object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 720–729.

[41] Z. Wu, S.-N. Lim, L. S. Davis, and T. Goldstein, "Making an invisibility cloak: Real world adversarial attacks on object detectors," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 1–17.

[42] 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.

[43] 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.

[44] Y. Cao, C. Xiao, D. Yang, J. Fang, R. Yang, M. Liu, and B. Li, "Adversarial objects against lidar-based autonomous driving systems," *arXiv preprint arXiv:1907.05418*, 2019.

[45] 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.

[46] 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.

[47] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham *et al.*, "Experimental security analysis of a modern automobile," in *2010 IEEE symposium on security and privacy*. IEEE, 2010, pp. 447–462.

[48] S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, S. Savage, K. Koscher, A. Czeskis, F. Roesner, and T. Kohno, "Comprehensive experimental analyses of automotive attack surfaces," in *20th USENIX security symposium (USENIX Security 11)*, 2011.

[49] W. Xu, K. Ma, W. Trappe, and Y. Zhang, "Jamming sensor networks: attack and defense strategies," *IEEE network*, vol. 20, no. 3, pp. 41–47, 2006.

[50] Y. W. Law, M. Palaniswami, L. V. Hoesel, J. Doumen, P. Hartel, and P. Havinga, "Energy-efficient link-layer jamming attacks against wireless sensor network mac protocols," *ACM Transactions on Sensor Networks (TOSN)*, vol. 5, no. 1, pp. 1–38, 2009.

[51] I. Rouf, R. Miller, H. Mustafa, T. Taylor, S. Oh, W. Xu, M. Gruteser, W. Trappe, and I. Seskar, "Security and privacy vulnerabilities of {In-Car} wireless networks: A tire pressure monitoring system case study," in *19th USENIX Security Symposium (USENIX Security 10)*, 2010.

[52] P. Ladisa, H. Plate, M. Martinez, and O. Barais, "Sok: Taxonomy of attacks on open-source software supply chains," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 1509–1526.

[53] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering–

a systematic literature review," *Information and software technology*, vol. 51, no. 1, pp. 7–15, 2009.
[54] A. Harzing, "Publish or perish," https://harzing.com/resources/publish-or-perish, 2023.
[55] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *international conference on machine learning*. PMLR, 2019, pp. 1310–1320.
[56] "Li-usb30-ar023zwdr," https://www.leopardimaging.com/product-category/usb30-cameras, 2023.
[57] I. Velodyne LiDAR, *VLP-16 User Manual*, https://velodynelidar.com/downloads/#datasheets, 2019.
[58] H. Lu, H. Zhang, S. Yang, and Z. Zheng, "Camera parameters auto-adjusting technique for robust robot vision," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 1518–1523.
[59] J. Tomasi, B. Wagstaff, S. L. Waslander, and J. Kelly, "Learned camera gain and exposure control for improved visual feature detection and matching," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2028–2035, 2021.
[60] "When to use your car's high-beam headlights: A complete guide," https://zutobi.com/us/driver-guides/when-use-high-beam-headlights, 2023.
[61] H. Wen, S. Chang, and L. Zhou, "Light projection-based physical-world vanishing attack against car detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
[62] A. Sayles, A. Hooda, M. Gupta, R. Chatterjee, and E. Fernandes, "Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 666–14 675.
[63] W. Zhu, X. Ji, Y. Cheng, S. Zhang, and W. Xu, "Tpatch: A triggered physical adversarial patch," 2023.
[64] T. Sato, Y. Hayakawa, R. Suzuki, Y. Shiiki, K. Yoshioka, and Q. A. Chen, "Revisiting lidar spoofing attack capabilities against object detection: Improvements, measurement, and new attack," *arXiv preprint arXiv:2303.10555*, 2023.
[65] B. Cyr, "Characterizing laser signal injection and its impact on the security of cyber-physical systems," Ph.D. dissertation, 2023.
[66] 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 ({USENIX} Security 20)*, 2020, pp. 877–894.
[67] 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 (USENIX Security 22)*, 2022, pp. 1903–1920.
[68] Z. Wang, Y. Wu, and Q. Niu, "Multi-sensor fusion in automated driving: A survey," *Ieee Access*, vol. 8, pp. 2847–2868, 2019.
[69] "Kitti benchmark," https://www.cvlibs.net/datasets/kitti/index.php, 2023.
[70] 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.
[71] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.
[72] H. Wu, C. Wen, S. Shi, X. Li, and C. Wang, "Virtual sparse convolution for multimodal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 653–21 662.
[73] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 35–52.
[74] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.
[75] S. Pang, D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 386–10 393.
[76] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1991–1999.
[77] G. Salton, "Modern information retrieval," *(No Title)*, 1983.
[78] D. Rukhovich, A. Vorontsova, and A. Konushin, "Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2397–2406.
[79] 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.
[80] 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.
[81] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[82] 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.
[83] J. Liu and J.-M. Park, ""seeing is not always believing": Detecting perception error attacks against autonomous vehicles," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2209–2223, 2021.
[84] A. Pfeuffer and K. Dietmayer, "Optimal sensor data fusion architecture for object detection in adverse weather conditions," in *2018 21st International Conference on Information Fusion (FUSION)*. IEEE, 2018, pp. 1–8.
[85] J. Kim, J. Choi, Y. Kim, J. Koh, C. C. Chung, and J. W. Choi, "Robust camera lidar sensor fusion via deep gated information fusion network," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1620–1625.
[86] W. Park, N. Liu, Q. A. Chen, and Z. M. Mao, "Sensor adversarial traits: Analyzing robustness of 3d object detection sensor fusion models," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 484–488.
[87] K. Yang, W.-Y. Lin, M. Barman, F. Condessa, and Z. Kolter, "Defending multimodal fusion models against single-source adversaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3340–3349.
[88] S. Wang, T. Wu, A. Chakrabarti, and Y. Vorobeychik, "Adversarial robustness of deep sensor fusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2387–2396.

Zizhi Jin received his B.S. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2020. He is currently working toward the PhD degree in the College of Electrical Engineering at Zhejiang University. His research interests include cyber-physical system (CPS) security, with a particular focus on autonomous driving security.

Xiaoyu Ji received his B.S. degree in Electronic Information & Technology and Instrumentation Science from Zhejiang University, Hangzhou, China, in 2010. He received his Ph.D. degree in Department of Computer Science from Hong Kong University of Science and Technology in 2015. From 2015 to 2016, he was a researcher at Huawei Future Networking Theory Lab in Hong Kong. He is now an associate professor with the Department of Electrical Engineering of Zhejiang University. His research interests include IoT security, including sensor, network, and AI security. He won the best paper award of ACM CCS 2017, ACM AsiaCCS 2018. He is a member of IEEE.

Yu Wang received his B.D degree in engineering from Zhejiang University, Hangzhou, China, in 2024. He will continue his doctoral studies in the College of Electrical Engineering at Zhejiang University. His research interests include Internet of Things(IoT) security.

Xuancun Lu received his B.S. degree in electrical engineering from Zheiiang University, Hangzhou, China, in 2023. He is currently working toward the PhD degree in the College of Electrical Engineering at Zhejiang University. His research interests include embodied AI security, with a focus on robotics security and autonomous driving security .

Yushi Cheng received her B.S. degree in Electrical Engineering from Zhejiang University in 2016, and her Ph.D. degree in Control Science and Engineering from Zhejiang University in 2021. She now is a postdoctoral researcher with the Department of Automation of Tsinghua University. Her research interests include IoT security, AI security, and mobile & ubiquitous computing. She received a WST best paper runner-up award in 2017, and an ASIACCS best paper award in 2018.

Chen Yan received the B.S. degree in Electrical Engineering in 2015 and the PhD degree in Control Theory and Engineering in 2021 from Zhejiang University, Hangzhou, China. He is currently an assistant professor at the College of Electrical Engineering, Zhejiang University. His research interests include sensing security, CPS security, and IoT security. He received the Best Paper Award of ACM CCS in 2017 and the Doctoral Dissertation Award of ACM China in 2021. He was acknowledged by Tesla Motors in the Security Researcher Hall of Fame in 2016.

Wenyuan Xu is currently a professor in the College of Electrical Engineering at Zhejiang University. She received her B.S. degree in Electrical Engineering from Zhejiang University in 1998, an M.S. degree in Computer Science and Engineering from Zhejiang University in 2001, and the Ph.D. degree in Electrical and Computer Engineering from Rutgers University in 2007. Her research interests include wireless networking, network security, and IoT security. Dr. Xu received the NSF Career Award in 2009, a CCS best paper award in 2017, and an ASIACCS best paper award in 2018. She was granted tenure (an associated professor) in the Department of Computer Science and Engineering at the University of South Carolina in the U.S. She has served on the technical program committees for several IEEE/ACM conferences on wireless networking and security, and she is an associated editor of TOSN.