Invisible but Detected: Physical Adversarial Shadow Attack and Defense on LiDAR Object Detection

Ryunosuke Kobayashi[1], Kazuki Nomoto[1,2], Yuna Tanaka[1], Go Tsuruoka[1], Tatsuya Mori[1,3,4]
[1]Waseda University, [2]Deloitte Tohmatsu Cyber LLC, [3]NICT, [4]RIKEN AIP

Abstract

This paper introduces "Shadow Hack," the first adversarial attack exploiting naturally occurring object shadows in LiDAR point clouds to target object detection models in autonomous vehicles. Shadow Hack manipulates these shadows, which implicitly influence object detection even though they are not included in output results. To create "Adversarial Shadows," we use materials that are difficult for LiDAR to measure accurately. We optimize the position and size of these shadows to maximize misclassification by point cloud-based object recognition models. Our evaluation is conducted on object detection models trained with the KITTI dataset, and the attack effectiveness is demonstrated within this setting. In simulations, Shadow Hack achieves a 100% attack success rate at distances between 11 m and 21 m across multiple models. Our physical world experiments validate these findings, demonstrating up to 100% success rate at 10 m against PointPillars and 98% against SECOND-IoU, using mirror sheets that achieve nearly 100% point cloud removal rate at distances from 1 to 14 meters. We also propose "BBValidator," a defense mechanism achieving a 100% success rate while maintaining high object detection accuracy. All data and code in this study are available at https://zenodo.org/records/14719074.

1 Introduction

As a key component of autonomous driving platforms, LiDAR sensors are essential for vehicles aiming for Level-4 autonomy and beyond. These advanced sensors use infrared laser pulses to create high-fidelity, three-dimensional maps of the vehicle's surroundings, enabling the real-time decision-making necessary for safe navigation. The integration of LiDAR technology has moved beyond the experimental stage, with many commercial robotaxi services now deploying LiDAR-equipped vehicles in urban areas around the world [2, 10]. This widespread adoption highlights the technology's essential role in translating theoretical autonomous capabilities into practical, real-world applications. However, as reliance on LiDAR systems increases, so does the potential impact of security vulnerabilities, raising concerns about the sensor data integrity and the safety implications for autonomous vehicle operations.

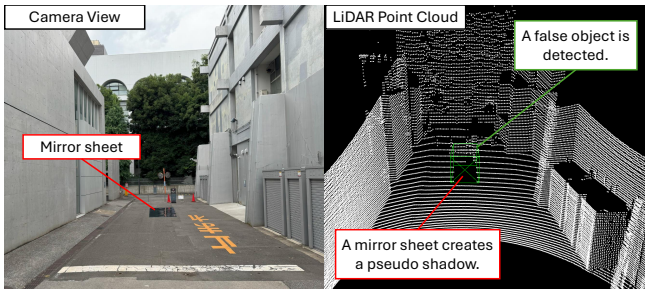

Figure 1: Shadow Hack overview: (Left) Camera view: A mirror sheet on the road. (Right) LiDAR point cloud: Pseudo shadow created by the mirror sheet, misclassified as a car (indicated by bounding box).

With the widespread adoption of autonomous vehicles and LiDAR technology, security vulnerabilities have become a critical focus for researchers and industry experts. Of particular concern are spoofing attacks on LiDAR sensors in autonomous vehicles [11, 14–16, 19, 26, 28, 29, 32, 33], which have emerged as a prominent threat. These attacks manipulate sensor readings by injecting malicious signals or strategically placing objects in the environment, leading to critical misclassifications by machine learning models. The consequences of successful spoofing can be severe, potentially causing autonomous vehicles to make dangerous decisions such as sudden braking, swerving, or failing to detect obstacles. This could result in collisions, traffic disruptions, or compromised passenger safety. By exploiting specific input patterns, attackers can induce false system outputs, undermining the reliability of autonomous driving systems.

This paper introduces "Shadow Hack," a novel attack vector against LiDAR-based sensing systems in autonomous vehicles. Unlike traditional spoofing attacks, Shadow Hack aims to manipulate the environmental context of LiDAR mea-

surements by strategically placing materials that are difficult for LiDAR to accurately measure, such as mirror sheets or infrared cut filters. This creates "Adversarial Shadows" within the LiDAR point cloud data, exploiting the naturally occurring shadows (occlusions) behind objects. The goal of Shadow Hack is to cause autonomous vehicles to misdetect non-existent objects, potentially triggering unnecessary emergency stops or continuous halting, which could lead to collisions or traffic congestion. By using flat objects placed during low-traffic periods, this method reduces the likelihood of detection or removal compared to previous attacks using 3D objects.

The key idea behind Shadow Hack is to exploit the naturally occurring "shadows" in LiDAR point cloud data. LiDAR sensors produce point cloud data representing the presence of objects, but this data also includes shadows cast behind objects (occlusion). Although object detection models do not explicitly output shadow information, as we demonstrate later, these shadows can influence the object detection process. Our method creates what we call "Adversarial Shadows" within the LiDAR point cloud using materials that are difficult for LiDAR to accurately measure, such as mirrored sheets, infrared cut filters, and infrared absorbing materials. By strategically placing these materials in the environment, we can create false shadows or gaps in the point cloud data, potentially causing object detection models to misinterpret the scene (See Figure 1). This approach represents a shift from active signal injection to passive environmental manipulation, presenting new challenges to existing security measures.

To design and evaluate the Shadow Hack, we take a comprehensive approach combining numerical optimization, realistic 3D simulation, and physical world experiments. We begin by optimizing the shape and size of the adversarial shadows to maximize the misdetection rate in point cloud-based object recognition models. These numerical experiments allow us to fine-tune the attack parameters for maximum effectiveness.

Our study evaluates Shadow Hack against three representative point-cloud object detection models: PointPillars [24], SECOND-IoU [31] (both voxel-based), and Point-RCNN [27] (point-based). Using AWSIM [20], an advanced autonomous driving simulator, we demonstrate the attack's effectiveness with nearly 100% success rates at distances between 11m and 21m. The attack shows high portability across different models and robustness under various environmental conditions, with success rates of approximately 100% in the 8 m to 23 m range for PointPillars and SECOND-IoU. Notably, PointPillars is used in real-world applications, such as the Apollo [1] open-source autonomous driving software, due to its real-time processing capabilities.

To validate Shadow Hack's real-world feasibility, we tested various materials that are difficult for LiDAR to accurately measure, finding mirrored sheets most effective with nearly 100% point cloud removal at 1-14 meters. Physical world experiments achieved a 100% success rate at 10 m and 98% at 15 m against PointPillars, with a 98% success rate at 10 m against SECOND-IoU. We also propose a defense mechanism, BBValidator, which maintains object detection accuracy while achieving a 100% defense success rate against Shadow Hack, demonstrating the potential for effective countermeasures.

We note that, as demonstrated through our experiments, Shadow Hack is effective against models trained on the KITTI dataset, a publicly available point cloud dataset widely used for object detection. As we show in Section 3, this dataset follows a strict annotation rule, meaning that annotations exist even when point clouds are absent due to occlusions. Such annotation practices make datasets more susceptible to the effects of Shadow Hack. In this study, we demonstrate the attack's effectiveness on models trained with the KITTI dataset, while its impact on models trained with other datasets that follow less strict annotation rules is discussed in Section 8.2

The key contributions of this work are:

- Introduction of Shadow Hack, a novel attack exploiting LiDAR point cloud shadows using materials difficult for LiDAR to measure accurately.
- Comprehensive material evaluation, demonstrating mirrored sheets' effectiveness in point cloud removal across multiple LiDAR types.
- Simulation-based validation of Shadow Hack, showing high success rates and transferability across different object detection models.
- Real-world experiment validation, confirming the attack's effectiveness against PointPillars and SECOND-IoU.
- Proposal of BBValidator, an effective defense mechanism against Shadow Hack with minimal impact on object detection accuracy.

2 Background and Related Work

2.1 LiDAR Mechanism

LiDAR emits laser pulses and measures their return time to calculate distance. A common type in autonomous vehicles is the mechanically rotating LiDAR, which provides a 360-degree view using multiple vertical lasers to create detailed 3D maps. This technology generates point cloud data with precise 3D positions, accurately representing the environment regardless of lighting conditions. However, as shown in the next section, there are scenarios where point cloud data may not be accurately measured despite this advanced mechanism.

2.2 Limitations of LiDAR Measurement

Despite its advanced mechanism, LiDAR technology has inherent limitations when measuring point clouds. This section outlines the conditions under which LiDAR sensors fail to accurately capture point clouds and highlights the technology's constraints. LiDAR sensors encounter measurement failures under three main conditions:

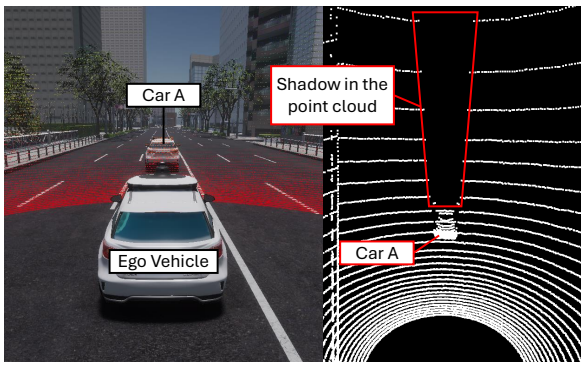

Figure 2: Illustration of point cloud shadow caused by 'Car A' obstructing LiDAR laser.

Occlusion by Other Objects. If multiple objects are present along the path of the laser pulse, only the object closest to the LiDAR is measured as a point cloud. Objects behind it are not measured and are detected as shadows, as shown in Figure 2. This phenomenon is also common with other light-based sensors such as cameras.

Laser Intensity Attenuation due to Distance. The point cloud measurement will fail if there are no objects in the path of the emitted laser pulse or if the distance between the object and the LiDAR sensor is too great. The maximum measurement distance varies by model and depends on factors such as laser wavelength, signal strength, and receiving sensor characteristics. For example, the maximum measurement distances of the LiDAR sensors used in this study are VLP-16: 100 m, OS1-64: 200 m, QT128: 50 m, and M1: 200 m.

Loss of Reflected Laser Light due to Material Properties. Certain material properties can prevent reflected light from returning to the LiDAR sensor, resulting in unmeasured point clouds. Examples include transparent glass, which causes total internal reflection, and mirrors, which reflect the laser light in a direction different from the angle of incidence. Similarly, infrared-absorbing fabrics absorb the laser light and prevent reflection.

The key idea of our attack is to exploit these limitations of LiDAR point cloud measurements to deceive object detection. Specifically, we focus on the loss of reflected laser light due to material properties. The attacker creates unmeasurable regions by placing materials such as transparent glass, mirrors, or infrared-absorbing fabrics on the road that the LiDAR sensor cannot accurately detect.

2.3 Object Detection Based on Point Clouds

Object detection based on point clouds is crucial for enabling autonomous driving systems to accurately perceive their surroundings. LiDAR provides precise 360-degree measurements, outperforming cameras in challenging conditions like darkness or adverse weather. As a result, many autonomous vehicles integrate cameras and LiDAR, leveraging the strengths of both sensors.

Object detection models using point clouds can be classified into two categories: point-based and voxel-based models. Point-based models, like PointRCNN [27], classify each point as an object or environmental point, detecting objects from the features of object points. Voxel-based models, such as PointPillars [24] and SECOND [31], divide the point cloud into 3D voxels and detect objects based on voxel features. While point-based models reduce false positives by processing each point individually, this method slows computation, making them less suitable for real-time detection in autonomous driving.

Autonomous driving systems use voxel-based models like PointPillars and CenterPoint for their fast processing and real-time object detection. Autoware [22] and Apollo [1], an open-source autonomous driving software, integrate these models into real-world vehicles. In this study, we specifically evaluate voxel-based point cloud detection models widely used in autonomous driving. To assess attack transferability, we examine the feasibility of the proposed attack across multiple models in Section 6.3.

2.4 Attacks against LiDAR-based Sensing

Attacks targeting perception and sensing systems based on LiDAR sensors in autonomous vehicles can be classified into two main categories: LiDAR Spoofing Attacks [14, 16, 19, 26, 28, 29] and Adversarial Object Attacks [11, 15, 32, 33]. In LiDAR Spoofing Attacks, attackers project laser beams that replicate real LiDAR reflections, causing the sensor to record fake point clouds. By timing the laser emissions, they can inject point clouds of any shape. In Adversarial Object Attacks, specially shaped objects are placed on roads or vehicles to trick the object detection model. These objects are designed to make the model either detect non-existent objects or miss real ones, leading to false positives or negatives and causing the vehicle to make incorrect decisions.

However, the deployability of these attacks is limited due to the need for expert-level laser emission skills and optimization of adversarial examples. The specialized equipment needed is often bulky and noticeable, making quick deployment in real-world situations difficult. LiDAR Spoofing Attacks necessitate injecting laser beams into a moving vehicle's LiDAR using equipment like drones or vehicles from a close range. Adversarial Object Attacks involve placing specially shaped objects on roads to deceive detection models.

In contrast, the attack proposed in this study is easier to deploy, using commonly available materials cut to specific sizes and placed on the ground. We note that a related study to our work is a recently presented work-in-progress paper [23], which explores adversarial shadow attacks as a preliminary study. Although the research direction of that work closely aligns with our own, our study differs in several key respects: we aim to achieve a robust attack against varying distances

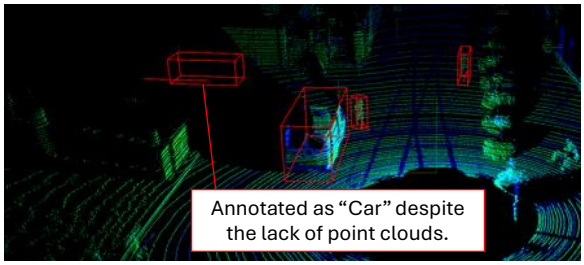

Figure 3: An example from KITTI dataset: ’Car’-annotated bounding box with no point cloud.

through shadow shape optimization, rigorously evaluate the attack’s robustness with respect to vehicle position and LiDAR models, conduct evaluations in physical environments, quantify the point cloud removal rate of our proposed shadow material, and newly develop and evaluate a defense method.

3 Preliminary Analysis

We analyze KITTI datasets [18] widely used for training point cloud object detection models as a preliminary step to the Shadow Hack attack. The dataset includes point clouds acquired by LiDAR, along with label data indicating the position, size, and class of objects such as Cars and Pedestrians within those point clouds. Ideally, annotations should be applied to point clouds that form the shape of the objects. However, there are instances in the dataset where areas are annotated as “Car” despite the absence of point clouds within the bounding box, as shown in Figure 3. This is likely because the annotators were instructed to annotate all objects, whether fully visible, partially occluded, or fully occluded, while viewing the simultaneously captured images and point clouds when annotating the KITTI datasets [18]. We focus on the number of points (N_{point}) within the bounding boxes that indicate the position and size of objects in the dataset as a clue to the proposed Shadow Hack. In this analysis, we counted the number of points within the bounding boxes for all objects annotated as “Car.”

Results. We found in the KITTI dataset, 332 out of 28,242 instances have $N = 0$ points, and 2,820 have $N \leq 10$ points (about 10% of total). Models trained on this dataset may erroneously detect objects even without point clouds. We propose an attack causing false detection of non-existent objects by controlling regions with no point clouds. As we will show later, false detections may be more likely when shadows are near objects, as these could be misinterpreted, especially in sparse point clouds.

4 Shadow Hack

4.1 Attack Overview and Threat Model

This study introduces “Shadow Hack,” an attack that exploits the vulnerability of object detection models in autonomous vehicles (AVs) to misinterpret regions without point clouds. The attack leverages the observation that these models can detect objects based solely on *shadows*, even when actual point clouds are minimal or absent. By strategically placing LiDAR-undetectable, optimized shapes on the ground, attackers can create artificial shadows that cause AVs to misdetect non-existent objects.

The primary goal of Shadow Hack is to disrupt the deep learning-powered object detection model integrated into the AV’s LiDAR system, causing it to falsely detect a non-existent vehicle. This misdetection can trigger an emergency stop or continuous halting, potentially leading to collisions with following vehicles or traffic congestion. Unlike previous attacks that used 3D objects on roadsides [32], Shadow Hack employs flat objects placed during low traffic periods, as shown in Figure 1. This approach reduces the likelihood of detection or removal and doesn’t require the attacker’s presence during execution.

Our threat model assumes an attacker with the following capabilities and limitations:

- Route Knowledge: The attacker has information about a specific point along the expected route of the target autonomous vehicle (AV). This knowledge could be obtained through observation of regular AV routes or publicly available information about autonomous taxi services.
- Model Access: The attacker has prior access to a point cloud object detection model similar to the one used by the target AV. This access does not require knowledge of the model’s internal structure or parameters. The attacker could obtain this by purchasing an AV of the same or similar model as the target, or by acquiring relevant design specifications or documentation, potentially through public sources or industry contacts.
- LiDAR Specifications: The attacker has knowledge of the LiDAR sensor specifications used in the target AV. This information could be obtained from manufacturer datasheets or through analysis of similar AV models.
- No Physical Access Required: Importantly, the attacker does not need physical access to the target AV or its internal systems. The attack can be prepared and executed without direct interaction with the vehicle.

These assumptions create a realistic yet challenging scenario. This model allows us to evaluate the potential real-world impact of the proposed attack method. It’s worth noting that if the attacker does not have a specific target in mind, the attack becomes even more feasible. In such cases, the

attacker can simply choose locations with high autonomous vehicle traffic for route knowledge. Similarly, they can focus on more popular object detection models and LiDAR specifications, which are likely to be used by a larger number of autonomous vehicles. This approach increases the likelihood of a successful attack without requiring specific knowledge about individual vehicles.

Under these assumptions, the attacker can manipulate the AV's perception system to trigger false detections and unnecessary emergency stops, presenting a significant security risk to autonomous driving systems. The potential for untargeted attacks further emphasizes the broad implications of this vulnerability for autonomous vehicle security.

4.2 Attack Framework

The Shadow Hack is a systematic approach designed to deceive autonomous vehicles by physically manipulating point cloud data. The framework consists of three distinct steps:

Step 1: Acquisition of Point Cloud Data. Initially, the attacker collects point cloud data (X_{benign}) in an environment with minimal surrounding objects. This data is gathered using a LiDAR sensor identical to that of the target vehicle to ensure data authenticity and maximize the attack's success rate. Data collection can be performed either in the real world or through simulation.

Step 2: Optimization of the Adversarial Shadow. Using the collected X_{benign} data, the attacker simulates and generates an adversarial shadow. This crucial phase involves optimizing the shadow's location and dimensions, represented by the size parameters (a, b, l). The objective is to manipulate the target vehicle's object recognition system into falsely detecting a non-existent object. Section 4.3 provides a detailed description of this optimization process.

Step 3: Real-World Implementation. The final step involves the practical deployment of the Adversarial Shadow. The attacker places Shadow Material of the optimized size (a, b, l) at the center of the lane where they intend the target vehicle to make an unexpected stop. These materials are specifically chosen for their LiDAR-undetectable properties, rendering them invisible in point cloud data and effectively creating a "shadow." As the target vehicle approaches the Shadow Material, it will erroneously detect a non-existent object, potentially triggering an unwarranted emergency response.

As we will reveal in Section 6, the Adversarial Shadow created in Step 2 demonstrates robust performance across various conditions/scenarios. As a result, potential attackers can effectively skip Steps 1 and 2, directly utilizing the optimized shadow parameters we have determined (Sec. 4.3). This significantly simplifies the attack process, making it more accessible and potentially more dangerous in real-world applications.

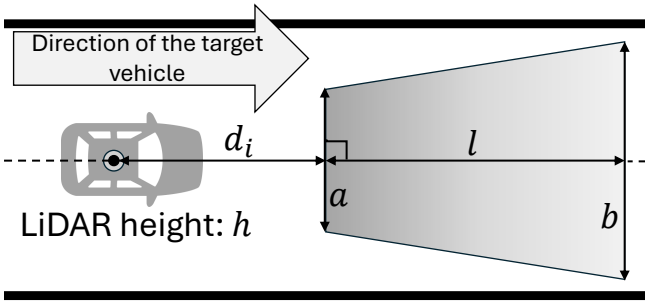

Figure 4: Trapezoidal model of Shadow Material with key dimensions relative to LiDAR.

4.3 Optimization of Adversarial Shadow

Methodology. We present a methodology for optimizing the size and shape of the Adversarial shadow. Our approach employs a realistic simulation that accurately models the Shadow Material and LiDAR scans in 3D space, allowing us to systematically evaluate various shadow configurations under controlled conditions. The optimization process is based on the results obtained by applying object detection machine learning models to the point clouds generated from these simulations. For a detailed description of the experimental setup, please refer to Section 6.1 and Fig. 10.

To optimize the Adversarial Shadow, we parameterize the shadow's dimensions based on observations of actual shadows in LiDAR point cloud data. We model the shadow as a trapezoid, which closely approximates the fan-like segments typically seen due to LiDAR's measurement mechanism. The trapezoidal model is defined by three key parameters: the length of the top base a, the length of the bottom base b, and the overall length l, as illustrated in Figure 4. This parameterization enables us to explore a range of shadow shapes and sizes efficiently.

Our optimization process aims to maximize the false positive rate, which we define as the number of frames in which non-existent objects are incorrectly detected at the locations where the Shadow Material is placed. To ensure robustness against varying distances between the LiDAR sensor and the Shadow Material, we evaluate the effectiveness of each shape configuration across a range of distances.

The optimization procedure involves measuring shadow point clouds at distances d_i ranging from 7 to 17 meters from the LiDAR sensor. This range was carefully chosen based on practical considerations. The lower bound of 7 meters corresponds to the closest point where the ground-hitting ray from an OS1-64 LiDAR sensor typically intersects when mounted at the standard height of a vehicle (1.73 m). The upper bound of 17 meters was selected because it is just within the critical braking distance for a vehicle traveling at 50 km/h, which is approximately 18 meters. By consistently detecting false objects within this range, we anticipate triggering maximum

braking control in autonomous vehicles. For each shape configuration and distance, we utilize an object detection model to compute a confidence score $C(d_i,a,b,l)$. We then calculate a cumulative score $S(a,b,l)$ for each shape by summing the instances where the confidence score exceeds a predetermined threshold τ; i.e., $S(a,b,l) = \sum_{i=1}^{n} \mathbb{I}(C(d_i,a,b,l) > \tau)$, where $\mathbb{I}$ is an indicator function that returns 1 when the condition is true and 0 otherwise, and n represents the number of distance samples. The optimal parameters (a,b,l) are obtained as $(a^*,b^*,l^*) = \arg\max_{a,b,l} S(a,b,l)$.

This approach ensures that the resulting Shadow Hack is not only highly effective in causing false detections but also robust to variations in distance from the LiDAR sensor, thereby enhancing its practical applicability in diverse scenarios.

Results. We begin by explaining our derivation of shadow length l^*. Preliminary experiments varying l showed that attack success rates decreased for lengths below 5m. Therefore, we adopted $l^* = 5$m as it represents the smallest Shadow Material size that maintains high attack efficacy. Details of these preliminary tests are omitted for brevity. For our main optimization, we set $l^* = 5$ m and the LiDAR height at 1.73 m, using an Ouster OS1-64 LiDAR sensor. Our target was a PointPillars model trained on the KITTI dataset, specifically using the pre-trained model distributed by OpenPCDet [30]. To ensure the shadow size remains within typical vehicle lane widths, we constrained our optimization to $1.0 \leq a \leq 4.0$ and $1.0 \leq b \leq 4.0$.

The optimization results reveal that the attack score S reaches its maximum value of 97 at two parameter settings: $(a^*,b^*,l^*) = (1.4,1.6,5.0)$ and $(2.8,2.4,5.0)$. This indicates that the object detection model becomes more susceptible to false detections when the values of a and b are close, suggesting that more rectangular Pseudo Shadows increase the attack success rate. This finding aligns with characteristics of the KITTI dataset used to train the object detection model. In this dataset, point cloud shadows labeled as "Car" tend to be rectangular, reflecting the typically box-shaped nature of actual vehicles and their resulting shadows. For optimal attack feasibility, we select the smallest shape among those with the highest score. Thus, this study adopts the shape $(a^*,b^*,l^*) = (1.4,1.6,5.0)$ for the Shadow Hack.

5 Performance of the Shadow Materials

This section evaluates the point cloud removal performance of various Shadow materials: mirror sheets, infrared cut films, and infrared-absorbing cloth.

Experimental Setup. Figure 6 shows the experimental setup. We measure the point cloud for 10 frames at each distance $d = 1,2,...,14$ m between the Shadow Materials and the LiDAR. Additionally, we record the point clouds obtained from asphalt, used as normal materials. For the point clouds measured over 10 frames, we analyze the number of points P_s measured on each Shadow Material and the number of points

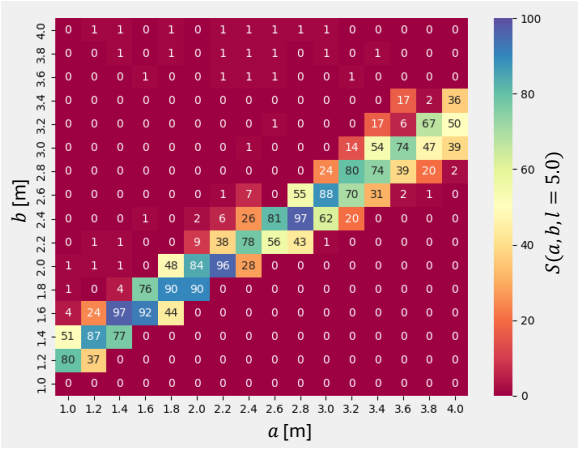

Figure 5: Heat map of optimization scores $S(a,b,l)$ for Shadow Hack's trapezoid shape parameters.

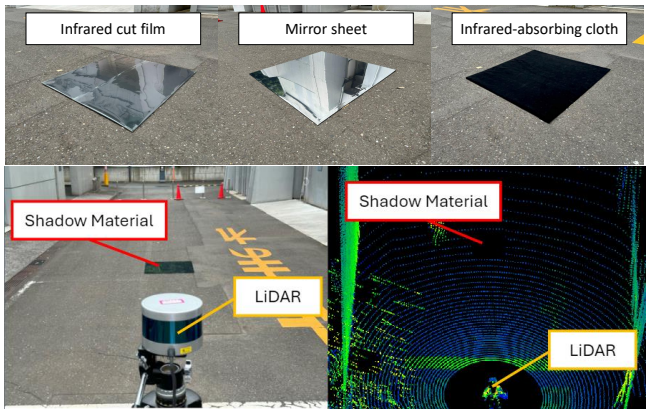

Figure 6: Top: Three Shadow Material types. Bottom: Experimental setup with Shadow Material on road center, scanned by LiDAR from distance d.

P_b measured on asphalt, as shown in Figure 7. We then calculate the point cloud removal rate due to the Shadow material as $1 - P_s/P_b$.

In the experiment, we used the Ouster VLP-16, OS0-64, OS1-64, Hesai QT128 and Robosense M1 as LiDAR sensors, as shown in Figure 1, all installed at a height of 1.73 meters. These LiDAR sensors are currently being equipped on autonomous vehicles under development worldwide. Among them, the M1 uses the MEMS method, while the others employ the rotating method, allowing evaluation of the two types of scanning methods. We selected mirror sheets, infrared cut films, and infrared-absorbing cloth as Shadow Materials. The mirror sheet reflects light emitted from the LiDAR in a different direction. The infrared cut film and infrared-absorbing cloth cut or absorb the light emitted by the LiDAR, significantly reducing the intensity of the reflected light returning to the LiDAR sensor.

Table 1: Specifications of the LiDAR used in this study. FOV: Field of View.

	VLP-16 [9]	QT128 [7]	OS0-64 [4]	OS1-64 [5]	VLS-128 [8]	Pandar 40P [6]	M1 [3]
Vertical Channel	16	128	64	64	128	40	-
Vertical FOV °	30	105.2	90	45	40	40	25
Horizontal FOV °	360	360	360	360	360	360	120
Wavelength [nm]	905	940	865	865	905	905	905
Max Range [m]	100	50	100	200	300	200	200
Min Range [m]	1	0.05	0.5	0.5	1	0.3	0.5
Scanning Type	Rotating	Rotating	Rotating	Rotating	Rotating	Rotating	MEMS

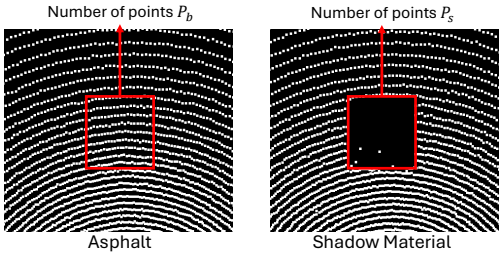

Figure 7: Top-view comparison of point clouds: asphalt (left) vs. Shadow Material (right). Point counts in red frames denoted as P_b and P_s respectively.

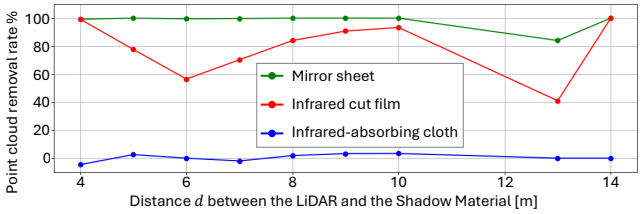

Figure 8: Point cloud removal rates of each Shadow Material for the OS1-64. For the 11 m and 12 m distances, the results were not plotted because Shadow Material fell between the LiDAR rays hitting the ground, resulting in zero measured point clouds for the Shadow Material.

Results. Figure 8 shows the point cloud removal rates of point clouds for different materials across various LiDAR sensors. The mirror sheet demonstrated the highest removal rate, making it ideal for Shadow Hack, with a 99.2% removal rate at 4 m and 100% at 14 m. The mirror sheet reflects laser light away from the incident angle due to total internal reflection, preventing the laser pulse from returning to the LiDAR sensor and causing the mirror to disappear from the point cloud. However, at 13 m, the removal rate dropped to 84.0%, likely due to surface irregularities in the handcrafted mirror sheet. The infrared cut film showed a 99.22% removal rate at 4 m and 93.3% at 10 m. The infrared cut film used in the experiment is a thermal insulation material for windows, specified to cut off 90% of infrared rays. Therefore, some of the light that was not fully cut off returned to the LiDAR, resulting in a lower point cloud removal rate compared to the mirror sheet.

Figure 9 shows the point cloud removal rates of point clouds for different materials across various LiDAR sensors. First, mirror sheet demonstrated an average point cloud removal rate of 99.8% for VLP-16, 98.3% for OS0-64, 98.2% for OS1-64, 96.5% for QT128, and 99.0% for M1. These results indicate that the mirror sheet achieves an approximate 100% point cloud removal rate, demonstrating its effectiveness not only against rotating LiDAR sensors but also against MEMS-based LiDAR sensors, regardless of the type of LiDAR sensor. This high effectiveness is due to the mirror sheet's property of simply reflecting light, unlike other materials. As a result, it is unaffected by variations in wavelength or other factors specific to different LiDAR sensors. This suggests that the mirror sheet is highly effective and optimal for executing attacks.

Second, we found that the point cloud removal rate of the infrared cut film varies depending on the type of LiDAR sensor. Specifically, the average point cloud removal rates for VLP-16, OS0-64, and M1 were 97.5%, 97.2%, and 88.5% respectively. In contrast, for the QT128, the removal rate was approximately 100% at distances of 5 to 7 m, but it frequently fell below 50% at other distances.

Finally, we found that the point cloud removal rate of the infrared-absorbing cloth is low for LiDAR models other than the VLP-16, making it unsuitable for attacks. For the VLP-16, the average point cloud removal rate was 99.5%. However, for the OS0-64, OS1-64, and QT128, the removal rate remained below 10% at most distances. In the experiment, the LiDAR sensor was installed parallel to the ground, projecting infrared laser pulses at an oblique angle to the material on the ground rather than at a right angle. Consequently, the cloth did not sufficiently absorb the infrared rays, leading to lower point cloud removal performance. The success of the attack against the VLP-16 can be attributed to the low sensitivity of its receiver or the weak intensity of its laser pulses.

6 Evaluation of Shadow Hack

This section presents a comprehensive evaluation of Shadow Hack's effectiveness across various scenarios and conditions.

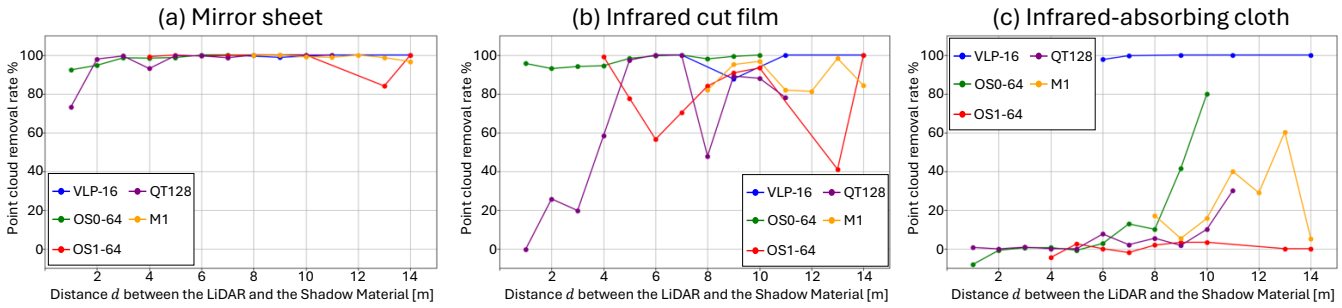

Figure 9: Point cloud removal rates for different materials across various LiDAR sensors. (a) Mirror sheet. (b) Infrared cut film. (c) Infrared-absorbing cloth. Sensors: VLP-16, OS0-64, OS1-64, QT128, M1. Similar to Figure 8, at distances where the Shadow Material fell between the LiDAR rays hitting the ground, resulting in zero measured point clouds, the results were not plotted.

We systematically assess the attack's performance with respect to the effective attack range (Sec.6.1), environmental settings (Sec.6.2), object detection models (Sec.6.3), and LiDAR sensor types (Sec.6.4). Additionally, we validate the feasibility of Shadow Hack in real-world conditions through physical experiments (Sec. 6.6).

6.1 Effective Attack Range

We evaluate the effective attack range by varying the distance d between the LiDAR-equipped vehicle and the Shadow Material. A wider effective range increases the risk of accidents in end-to-end autonomous driving systems.

Experimental Setup. We utilized AWSIM, a simulator for developing and evaluating the autonomous driving open-source software Autoware, to collect point clouds. AWSIM realistically simulates LiDAR sensors by adding Gaussian noise to the LiDAR point clouds, thereby replicating the characteristics of real-world LiDAR sensors. We used the Ouster OS1-64 LiDAR, positioned at a height of 1.73 meters above the ground, because its Vertical FOV and Vertical channel are similar to those of the HDL-64 LiDAR used in the KITTI dataset. The measurement locations (scene 1–5) were randomly selected from five scenes within the default map provided by AWSIM, which faithfully recreates Tokyo.

Figure 10 illustrates the experimental setup. Within the AWSIM simulation environment, we faithfully reproduced the characteristics of the mirror sheet, which had the highest point cloud removal rate among the Shadow Material described in Section 5. As shown in Figure 9-(a), the mirror sheet achieved approximately 100% point cloud removal rate at almost all distances for all LiDARs. Therefore, in AWSIM, we implemented the Shadow Material as an ideal Shadow Material that does not appear in the point cloud at all.

We used OpenPCDet [30] for 3D object detection toolbox in point clouds. OpenPCDet provides a variety of pre-trained models. We focused our evaluation on PointPillars, a Voxel-based feature extraction model trained on the KITTI dataset.

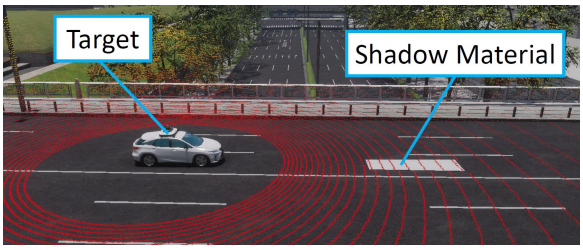

Figure 10: Experimental setup in AWSIM. The Shadow Material is placed in the middle of the target vehicle's lane, and point cloud data is collected as the vehicle approaches. Reflections from the Shadow Material are excluded to replicate its characteristics.

The detection threshold is set to the default value of 0.1. OpenPCDet also includes other pre-trained models.

Results We obtained 10 frames of point cloud data at each distance $d = 1, 2, ..., 30$ m between the vehicle equipped with a LiDAR sensor and the Shadow Material in scene 1, and performed object detection inference on each frame. We analyze the inference results of each frame and calculate the object detection rate on the Shadow Material at each distance. Since nothing is placed on the Shadow Material, the expected object detection rate should be zero if the attack is non-feasible. However, if objects are incorrectly detected on the Shadow Material, it is considered a successful attack, and the Attack Success Rate is calculated.

Figure 11 shows the results. The results indicate that in Scene 1, the attack was continuously successful in the range of $11\text{m} \leq d \leq 21\text{m}$. Additionally, it was observed that the autonomous vehicle suddenly detected an object at $d = 21$ m, which had not been detected at distances greater than 23m. Considering an autonomous vehicle traveling at a speed of 60 km/h necessitates a braking distance of 27 m, the detection of an object at a distance of 21 m, falling short of the required braking distance, would necessitate the initiation of emergency braking procedures. There exists a significant risk of the following vehicle colliding with the autonomous ve-

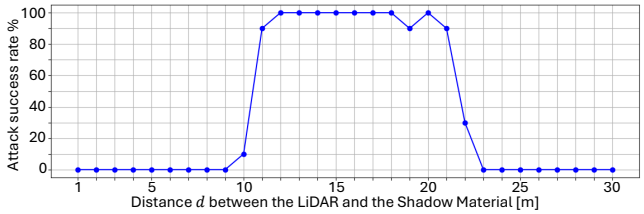

Figure 11: Attack success rate for varying distance d.

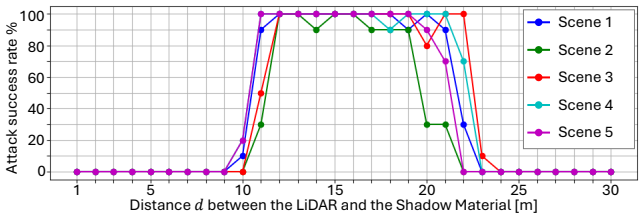

Figure 12: Attack success rate across Scenes 1–5.

hicle when the distance between them is minimal, or when the braking performance of the following vehicle, such as large buses or trucks, is inferior to that of the autonomous vehicle. We note that no objects were detected at distances of 9 m or less. As shown in Figure 10, this can be attributed to the autonomous vehicle's inability to observe the ground at very close distances, preventing the Shadow from appearing in the point cloud as intended. Furthermore, as discussed in Section 4.3, the optimized attack distance ranged between 7 m and 17 m, further supporting this observation.

6.2 Surrounding Environments

We evaluate the robustness of the attack against diverse surrounding environments by expanding our simulation to include four additional autonomous driving scenes, resulting in a total of five distinct scenarios (Scenes 1–5).

Experimental Setup. We conducted experiments in five different scenes, Scene 1 – 5, as shown in Appendix A.1. We place the Shadow Material, optimized in Section 4.3, and the LiDAR sensor in each of Scenes 1–5. In each scene, we obtained 10 frames of point cloud data at each distance d, following the procedure described in Section 6.1, and performed object detection inference on each frame. We analyzed the inference results and evaluated the object detection rate on the Shadow Material at each distance in each scene. The target object detection model used is PointPillars, and the LiDAR sensor used is the Ouster OS1-64, as described in Section 6.1.

Results. Figure 12 shows the experimental results, demonstrating that the Shadow Hack is robust to changes in the surrounding environment. These results indicate that the attack is feasible at distances between 12 m and 19 m in all scenes. In Scene 2, the attack success rate exceeds 90% for distances between 12 m and 19 m. In Scene 3, the success rate is nearly 100% for distances between 12 m and 22 m. In Scene 4, the success rate is nearly 100% for distances between 11 m and 21 m. In Scene 5, the success rate exceeds 90% for distances between 11 m and 20 m.

6.3 Object Detection Models

We evaluate the robustness of the Shadow Hack on various object detection models.

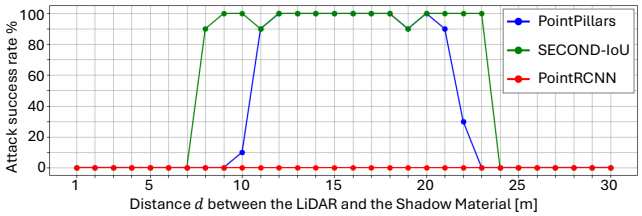

Figure 13: Attack success rate across three object detection models.

Experimental Setup. We conducted experiments using three prominent 3D object detection models: PointPillars, SECOND-IoU, and PointRCNN. Each model was trained using the KITTI dataset. The feature extraction mechanisms differ among the models. PointPillars and SECOND-IoU use voxel-based feature extraction, while PointRCNN uses point-based feature extraction. The object detection thresholds for each model are set to their default values: 0.1 for PointPillars, SECOND-IoU, and PointRCNN. Following the methodology in Section 6.1, we input point cloud data acquired from various distances into each model and analyze the results.

Results. Figure 13 shows the experiment results. These results indicate that the proposed attack is robust against the voxel-based object detection model. The voxel-based object detection model, SECOND-IoU, achieves an attack success rate of approximately 100% for distances between 8 m and 23 m. PointPillars and SECOND-IoU are both voxel-based object detection models. PointPillars partitions point cloud data into pillars, while SECOND-IoU partitions it into voxels. These models detect objects based on the features of each pillar or voxel, including shadows, rather than on the raw point cloud. Pseudo Shadows, which are regions without point clouds, affect the features of these pillars and voxels, leading to false detections. On the other hand, the attack success rate for the point-based object detection model PointRCNN is 0% regardless of d. Point-based object detection models detect objects using features of point clouds classified as objects, excluding shadows. Therefore, even if points are dropped due to the attack, it does not affect the object detection process.

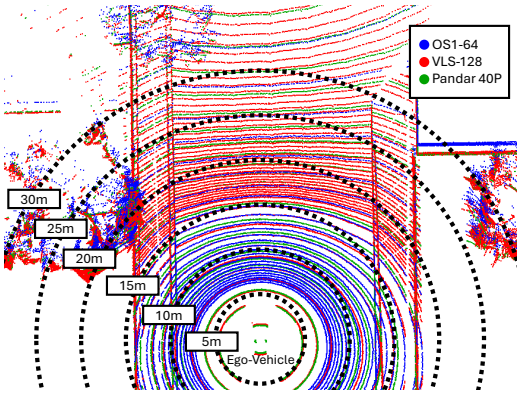

Figure 14: An overhead view of the point clouds measured at the same location using OS1-64, VLS-128, and Pandar 40P. This demonstrates the varying density of LiDAR rays depending on the distance.

6.4 LiDAR Sensor Models

We evaluate the robustness of attacks using different LiDAR sensor models from those used in the optimization process.

Experimental Setup. We place shadows optimized for the OS1-64 in Section 4.3 in front of various LiDAR sensors. The evaluated LiDAR sensors include OS1-64, VLS-128, and Pandar 40P, each possessing distinct characteristics. Table 1 summarizes the features of these LiDARs, while Figure 14 illustrates the point clouds collected from the same location by each LiDAR. The OS1-64 has 64 vertical laser beams, with the spacing between the beams on the ground increasing as the distance from the LiDAR grows. Velodyne's VLS-128 has 128 vertical laser beams, with significant variation in beam spacing based on distance. Specifically, beam spacing is wider up to 15 meters, becomes denser beyond that, and then sparse again at greater distances. The Pandar 40P, with its 40 vertical laser beams, maintains uniform beam spacing compared to other LiDARs, resulting in consistent ground coverage.

Results. Figure 15 presents the experimental results. These results indicate that, for all LiDAR sensors, the object detect model mistakenly identifies the Shadow as an object. This demonstrates the high robustness of the proposed attack against LiDAR sensors. Furthermore, it is evident that the effective attack range varies depending on the type of LiDAR sensor used. For the OS1-64 sensor, the attack success rate is approximately 100% when the distance d ranges from 11 m to 21 m. For the VLS-128 sensor, the attack success rate is around 100% for distances ranging from 4 m to 12 m and from 21 m to 30 m. Similarly, for the Pandar 40P sensor, the attack success rate is approximately 100% when the distance d ranges from 11 m to 27 m.

The variation in attack success distances across different LiDAR types can be attributed to the differences in the spacing of laser beams hitting the ground. Figure 16 shows the distribution of laser beams hitting the ground for each LiDAR and the locations for successful placement of attack shadow. In the experiments, the OS1-64 LiDAR model was targeted, and shadows optimized for successful attacks were used under conditions where the spacing between ground-hitting rays ranged from 11 to 21 m. When targeting other LiDAR models, attacks succeed if the spacing between ground-hitting rays matches the conditions mentioned above; if not, the attacks fail. With the VLS-128 LiDAR model, attacks are successful when the distance d ranges from 4 to 12 m, as the spacing between rays is 1.12 m. Conversely, at distance d between 14 and 18 m, the spacing falls below 0.57 meters, which is outside the optimized range, leading to attack failure.

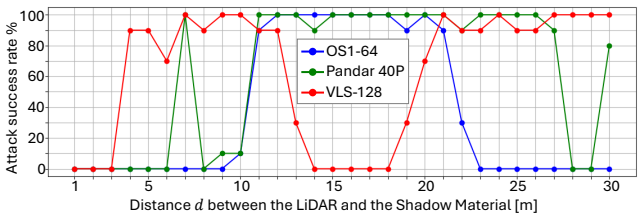

Figure 15: Attack success rate across three LiDAR sensors.

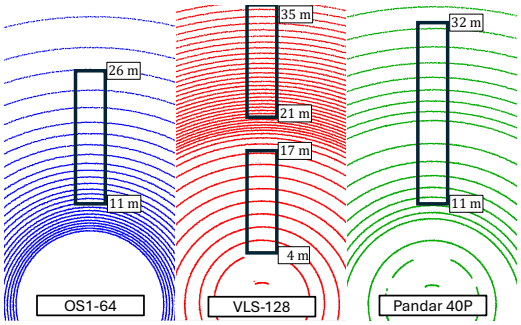

Figure 16: The distribution of laser beams hitting the ground for each LiDAR and the locations for successful placement of attack shadows (areas enclosed by rectangles in the figure). The distances in the figure are from the LiDAR sensor.

6.5 LiDAR Location

We evaluate the robustness of the attack against the location of the LiDAR and Shadow Material by shifting the LiDAR horizontally and vertically.

Experimental Setup. We obtained 10 point cloud frames at each distance d in Scene 1, following Section 6.1, and performed object detection on each frame. In the same environment as Section 6.1, we also evaluated two scenarios: (1) horizontal shifts, where the vehicle was offset by 1.5m from the Shadow Material lane, and (2) a height shift, where the LiDAR was mounted 30cm higher than in Section 6.1 while driving in the center of the lane.

Results. The experiments revealed that Shadow Hack is a robust attack against the positioning of the LiDAR and the

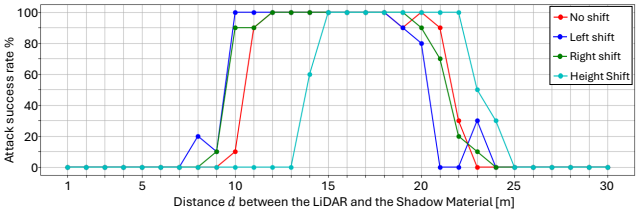

Figure 17: Attack success rate across different positions.

Shadow Material. Figure 17 presents the experimental results, demonstrating Shadow Hack's resilience to variations in the spatial relationship between the LiDAR and the Shadow Material. (1) In the left and right shift scenarios, the attack success rate exceeded 90% between 9 m and 19 m for the left shift and between 9 m and 20 m for the right shift. Compared to the no-shift scenario, where the attack success rate exceeded 90% from 11 m to 21 m, the success range in the shifted scenarios remained nearly identical. (2) In the height shift scenario, the attack achieved a 100% success rate between 15 m and 22 m. Although the success range shifted farther from the Shadow Material compared to the no-shift scenario, the attack remained highly effective. The increased attack success range in the height shift scenario likely resulted from the elevated placement of the LiDAR, which caused its rays to reach the ground at positions farther from the LiDAR itself. Consequently, the ray spacing matched the range described in Section 6.4 at greater distances from the LiDAR overall.

6.6 Physical-World Attack

We evaluate the feasibility of the Shadow Hack in the physical-world.

Experimental Setup. Figure 18 illustrates the experimental setup. The optimized Shadow Material is placed at the center of the road. The LiDAR is positioned at distances of d = 5m, 10m, 15m, 20m, and 25m from the Shadow Material, and point clouds are collected at each distance. We adopted a mirror sheet with a reflectivity of 98% for the Shadow Material, as it demonstrated the highest point cloud removal performance in section 5. The LiDAR used is an Ouster OS1-64, installed at a height of 1.73 m from the ground. In the experiment, two object detection models, PointPillars and SECOND-IoU, which were confirmed to be vulnerable to attack in the simulation experiments in Section 6.3, are selected as attack targets. These models were trained on the KITTI dataset and are provided by the OpenPCDet framework [30].

Results. We analyzed the object detection results for the point clouds collected at each distance d between the LiDAR sensor and the Shadow Material. We collected 50 frames of point clouds at each distance d to account for the vibration of the measured point clouds caused by LiDAR noise, and performed object detection inference on each frame. We analyzed the detection results and evaluated the object detection

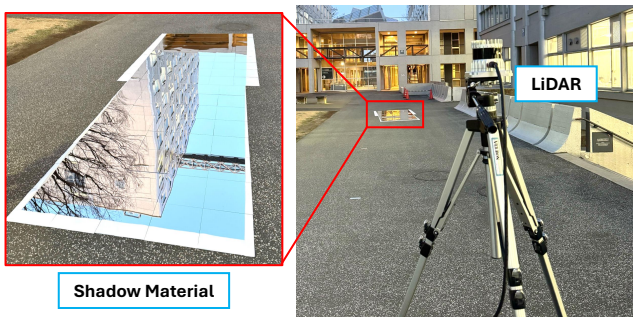

Figure 18: Real-world attack experiment setup. A mirror sheet was placed on the road as Shadow Material.

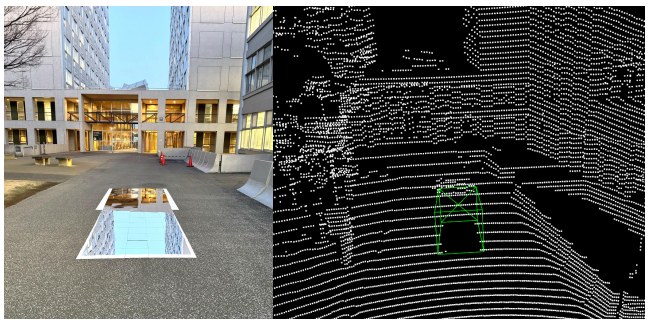

Figure 19: Left: No objects on the Shadow Material (actual). Right: False car detection in LiDAR point cloud.

rates. As there is nothing placed on the Shadow Material, the object detection rate should ideally be zero.

Table 2 presents the results of the attack success rates, while Figure 19 illustrates the object detection results when the attack is successful at a distance of 10 m. These results demonstrate that both PointPillars and SECOND-IoU are capable of achieving high attack success rates in physical environments, underscoring the feasibility of such attacks in real-world scenarios. PointPillars achieves attack success rates of 100% at a distance of 10m and 98% at 15 m, showcasing its effectiveness within this range. SECOND-IoU, on the other hand, achieves a 100% attack success rate at 10m but does not perform as effectively beyond this range. This difference can be attributed to the models' robustness to physical-world noise. PointPillars maintains high attack success rates despite noise, while SECOND-IoU, though effective in simulations, is limited to shorter ranges under noisy real-world conditions.

These findings highlight the significant potential of such attacks to disrupt autonomous vehicle operations, particularly at close distances, where there is a high likelihood of triggering critical control actions, such as sudden braking, potentially leading to accidents.

Table 2: Average attack success rate by distance for each model in the real world experiments.

Model	Distance				
	5m	10m	15m	20m	25m
PointPillars	0%	**100%**	**98%**	0%	0%
SECOND-iou	0%	**98%**	0%	12%	0%

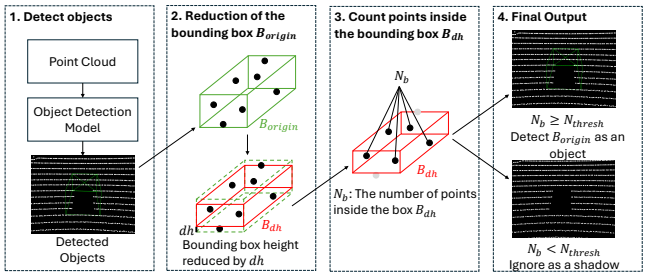

Figure 20: Overview of BBValidator: (1) Object detection, (2) Bounding box reduction by dh, (3) Point counting within reduced box, N_b, (4) Classification based on N_b and N_{thresh}.

7 Defense

We develop a defense mechanism BBValidator against Shadow Hack and evaluate its effectiveness against the attacks as well as its impact on non-attack object detection accuracy as a side effect.

7.1 Methodology

BBValidator disables attacks by validating object detection results using the number of point clouds within bounding boxes. Figure 20 shows an overview of the validation mechanism. The object detection model outputs bounding boxes that contain no point clouds when false detections are triggered by attacks. The validation mechanism counts the number of point clouds within the bounding boxes output by the object detection model. The validation mechanism is determined by two parameters, N_{thresh} and dh. The detection result is identified as a false detection when the counted number of point clouds is below a certain threshold N_{thresh}. To accurately count the number of point clouds corresponding to objects, simply counting the point clouds within the detected bounding boxes may include ground points, leading to inaccuracies. Therefore, we count the number of point clouds within a bounding box whose bottom surface has been moved upwards by dh from the original bounding box.

7.2 Evaluation of Defense Effectiveness

We evaluate the BBValidator's defense success rate against Shadow Hack attacks. The evaluation is performed across various combinations of its two key parameters: N_{thresh} and dh. In this experiment, 111 frames where attacks were successful, collected in Section 6, are input into the BBValidator-integrated PointPillars object detection model. We evaluate the effectiveness of our proposed defense mechanism by analyzing the output of the BBValidator-integrated PointPillars. We judge a frame as a successful defense if no false positive bounding boxes appear on Shadow Material, and we calculate the defense success rate accordingly. The parameters of the BBValidator in the experiment are $N_{\text{thresh}} = 1, 2, ..., 20$, $dh = 0.00, 0.05, ..., 0.20$.

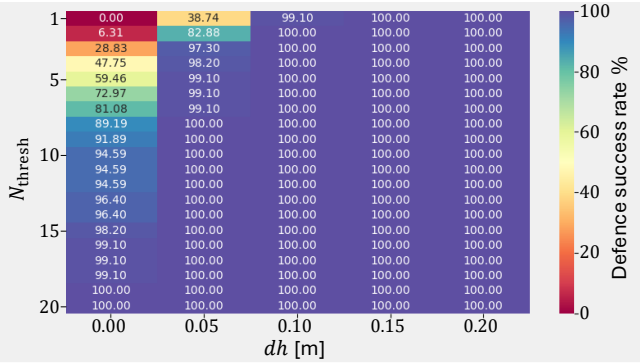

Figure 21: Heat map of defense success rates across (N_{thresh}, dh) parameter space.

Figure 21 shows the defense success rates for various parameters of the BBValidator. These findings indicate that by appropriately configuring the parameters (N_{thresh}, dh), the defense success rate can achieve 100%, thus confirming the effectiveness of BBValidator against attacks. Specifically, a defense success rate of 100% was observed when $dh = 0.0$ and $N_{\text{thresh}} \geq 19$, as well as when $dh = 0.05$ and $N_{\text{thresh}} \geq 8$. When dh is set to a small value, the BBValidator validation mechanism counts ground points as part of the target object's point cloud, which tends to result in a higher N_{thresh} being required for effective defense. Conversely, when N_{thresh} takes on a high value, there is an increased likelihood of erroneously removing results from non-attacked bounding boxes. Section 7.3 evaluates the side effects of the proposed defense method on object detection accuracy in non-attack scenarios.

7.3 Evaluation of Defense Side Effects

We investigate potential side effects of the BBValidator by assessing its impact on object detection accuracy under non-attack conditions. We evaluate the object detection accuracy on the KITTI dataset for each parameter set (N_{thresh}, dh) of the BBValidator. We prepare BBValidator-integrated PointPillars with each parameter set (N_{thresh}, dh) Next, we input point clouds from the validation set of the KITTI dataset into the BBValidator-integrated PointPillars models and measure the output results of these models. Finally, we calculate the 3D

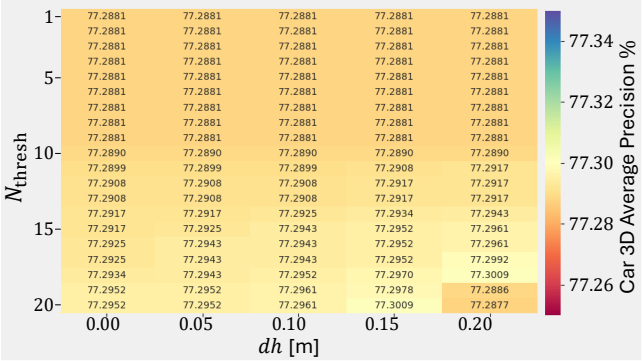

Figure 22: Heat map of 3D Average Precision on KITTI dataset across (N_{thresh}, dh) parameter space. Baseline 3D AP without BBValidator: 77.2881.

Average Precision (3D AP), a commonly used metric for evaluating LiDAR object detection models, to assess the impact of the BBValidator parameters (N_{thresh}, dh) on the detection accuracy under non-attack conditions. Additionally, we perform the aforementioned analysis using the normal PointPillars model as a baseline for comparison.

The results demonstrate that integrating the BBValidator into the PointPillars model has no significant impact on detection accuracy under non-attack conditions. Figure 22 illustrates the results of the experiments. The change in detection accuracy due to the integration of the defense mechanism was within approximately 0.02 for all parameter. Additionally, the 3D AP before implementing the defense mechanism was 77.2881. The change in detection accuracy due to the integration of the defense mechanism ranged from -0.0004 to +0.0128. The 3D AP decreased by -0.0004 compared to before the implementation of the defense mechanism only when $(N_{\text{thresh}}, dh) = (20, 0.2)$. In other cases, the integration of the defense mechanism resulted in a slight increase in 3D AP. We attribute this improvement to the BBValidator's contribution in preventing false-positive detections.

8 Discussion

8.1 Challenges in the Real-World Scenarios

Stealth of the Attack. The Shadow Hack involves placing flat materials that LiDAR cannot detect in the middle of the road. To be successful, an attacker must place these materials stealthily. Optimal attack locations include areas with low vehicle and pedestrian traffic or during nighttime hours. While it's theoretically possible to deploy the Shadow Material from a moving vehicle, such as a truck in front of the target autonomous vehicle, precise placement at the correct location and angle would be challenging. In addition, the attacker must ensure that the Shadow Material remains in place until the attack is executed to avoid removal by suspicious third parties. The flat shape of the Shadow Material allows regular vehicles to drive over it without disturbance. In addition, as shown in Figure 19, mirrored sheets can resemble puddles of water from a distance, reducing the likelihood of removal.

Durability of the Shadow Material. The durability of Shadow Materials presents a significant challenge due to their inherent thinness, as demonstrated in our real-world attack assessment Section 6.6), where we reinforced mirror sheets with plastic corrugated cardboard. However, this solution has limitations: the specular reflection critical for point cloud removal can be compromised by physical deformation, surface damage, or contamination, while the plastic boards remain susceptible to bending under the weight of the vehicle. More robust materials, such as wood or metal plates, could improve durability and attackability. However, we also note that environmental factors still pose significant obstacles. Rain can degrade performance by accumulating water droplets as shown in Appendix A.2, and strong winds risk displacing the material altogether. These practical limitations underscore the need for careful planning and consideration of weather conditions when using Shadow Hack in real-world scenarios, potentially limiting the window of opportunity for attacks and highlighting the delicate balance between effectiveness and practicality when executing this type of exploit.

8.2 Target Model

Target LiDAR model. Our simulation evaluated attack success rates against different LiDAR devices as shown in Figure 15. The results suggest a relationship between the spacing of LiDAR beams hitting the ground and attack success rates that varies between LiDAR models. This variation influences the effective distance between the Shadow Material and the LiDAR to induce false positives. As a result, Shadow Material placement must take into account both the desired deceleration location and the attackable range for each specific LiDAR model. LiDAR sensors with very few vertical channels (e.g., 16 for Velodyne VLP-16) may have insufficient ground-hitting rays to accurately reproduce the pseudo-shadow shape in the point cloud. Our results indicate that the attack is most effective against LiDAR models with 40 or more vertical channels, such as the three LiDAR devices used in our study. Additionally, focusing on the LiDAR scanning mechanisms, all the LiDAR devices evaluated in Section 6.4 utilize the rotating scanning method. Evaluating MEMS-based or solid-state LiDARs requires training models on datasets specifically collected using those scanning methods. However, publicly available datasets generally lack such data, presenting a challenge for future research. Nevertheless, as demonstrated in Section 5, the mirror sheet successfully removed point clouds from the MEMS-based LiDAR device M1. This result suggests that our attack method holds promise for success against MEMS-based LiDARs as well.

Target Object Detection Model. Our evaluation of attack

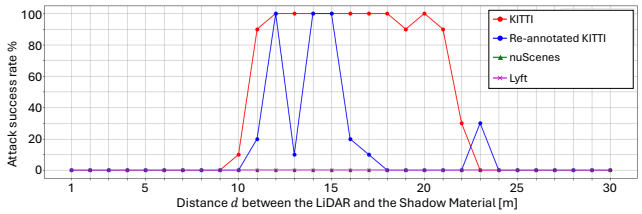

Figure 23: Attack success rate based on the dataset used for training the PointPillars model.

success rates (see Figure 13) showed that the Shadow Hack was effective against voxel-based models like PointPillars and SECOND-IoU, but not against the point-based PointRCNN model. There are distinct roles and strengths for point-based and voxel-based object detection models. Firstly, point-based models process LiDAR point clouds directly, capturing fine details and spatial relationships, which makes them excellent for high-precision detection but computationally expensive and unsuitable for real-time processing. On the other hand, voxel-based models, on the other hand, convert point clouds into a 3D grid (voxels), reducing computational costs and improving real-time performance. This makes voxel-based models ideal for scenarios requiring efficient real-time processing or when resources are limited. In autonomous vehicles, where real-time performance is crucial, voxel-based models are often preferred. Each of these models has specific advantages and applications. In the context of autonomous vehicles, where real-time performance is critical, voxel-based models tend to be preferred. Therefore, the Shadow Hack proposed in this study is effective against voxel-based object detection models and has a significant impact on autonomous driving systems.

Impact of Annotation Rules of Point Cloud Datasets. As shown in Section 3, the KITTI dataset employs stringent annotation rules that label objects even when no LiDAR points are detected, facilitating continuous tracking of briefly disappearing objects. In contrast, other datasets such as nuScenes [13] and Lyft [17] label only objects that contain LiDAR points, making its annotation rules less stringent. This difference may impact the model vulnerability to adversarial attacks.

We have revealed that the Shadow Hack is successful on models trained with the KITTI dataset, which employs stringent annotation rules. We conducted additional experiments using object detection models trained on the nuScenes dataset and the Lyft dataset (Figure 23) and found that the attack is not successful in this case. To further assess the generality of the observations, we re-annotated the KITTI dataset using the similar annotation rule to nuScenes and evaluated the attack (Figure 23), confirming that it becomes significantly less effective. These findings suggest that Shadow Hack is considered effective only on models trained on strictly annotated point cloud datasets. Evaluating its effectiveness on models trained with other datasets that follow a variety of annotation rules remains a topic for future work.

We note that stringent rules are essential for labeling fully occluded objects (potentially with 0% visibility), as discussed in the nuScenes forum [25]. Such rules ensure robust object tracking in autonomous vehicle applications, where continuous, reliable tracking is crucial. Consequently, while relaxed annotations mitigate the attack, the practical need for stringent labeling keeps the Shadow Hack attack a significant concern.

8.3 Multi Sensor Fusion

In general, autonomous vehicles perform object detection by combining multiple sensors, such as cameras and LiDAR. Among the various approaches to multi sensor fusion (MSF), this work focuses on a widely adopted method: performing perception independently with each sensor and integrating their outputs, as highlighted in [19]. In the following, we discuss attacks and defenses related to this type of MSF model. We adopted an implementation using open-source software that allows replication of results. Specifically, we adopted PointPillars for point cloud-based object detection and YOLOv5 [21] for image-based object detection, and the model detects an object if either modality reports a detection.

Shadow Hack Evaluation on MSF. To evaluate the Shadow Hack against MSF, we used the same experimental setup as described in Section 6.1. While approaching the Shadow Material, we collected point cloud data from consecutive frames and corresponding onboard camera images, which were then input into the MSF model. The results in Appendix A.3 showed that while the camera images failed to detect anything even in the presence of the Shadow Material, the point cloud data falsely detected nonexistent vehicles, similar to findings in Section 6.1. This demonstrates that the Shadow Hack attack is effective against MSF systems.

Integrating BBValidator with MSF. We also evaluated whether combining BBValidator with MSF could mitigate false negatives. The evaluation involved two datasets: *Attacked data*, collected while approaching the Shadow Material, and *Non-attacked data*, collected while approaching actual vehicles. BBValidator was implemented within the PointPillars module of the MSF model. The results indicated that during the attack, BBValidator successfully prevented false detections, with no false detections occurring in either the images or the point clouds. In non-attack scenarios, BBValidator did not introduce any false detections. However, increasing the threshold value N_{thresh} occasionally caused the point cloud module to misdetect actual vehicles as False Negatives. Despite this, the image module consistently detected the vehicles, allowing the MSF model as a whole to correctly identify them. This highlights the robustness of combining BBValidator with MSF for reliable detection in both attacked and non-attacked scenarios.

9 Conclusion

We introduce Shadow Hack, the first attack that exploits shadow characteristics in LiDAR systems. This approach manipulates LiDAR point clouds by strategically removing points with materials that are difficult for LiDAR to measure accurately, a novel concept in autonomous vehicle security. Shadow Hack creates pseudo-shadows by placing optimized Shadow Materials on road surfaces, deceiving object detection models without interfering with LiDAR hardware or data streams. Our evaluation showed that mirror sheets were especially effective, with a 98% point cloud removal rate at distances from 1 to 14 meters. Physical experiments and 3D simulations confirmed the real-world feasibility of Shadow Hack, demonstrating high success rates against models like PointPillars and SECOND-IoU across various distances and conditions. However, our results suggest that Shadow Hack is effective only on the strictly annotated KITTI dataset, where annotations exist despite the absence of point clouds. Evaluating its effectiveness on models trained with other strictly annotated datasets remains a topic for future work. Our analysis of the attack's effectiveness across different LiDAR models revealed valuable insights into its adaptability and limitations. To address the security risks posed by Shadow Hack, we proposed a defense mechanism called BBValidator, which achieved a 100% defense success rate with minimal impact on detection accuracy. This study identifies a key vulnerability in LiDAR-based systems and offers a framework for mitigating such attacks. By demonstrating both an effective attack and defense, we contribute to enhancing the robustness of autonomous vehicles. Our findings highlight the value of multi-layered security approaches, combining sensor improvements, advanced detection algorithms, and specialized defenses. This research may guide future efforts to secure LiDAR systems against attacks exploiting object concealment or obfuscation.

Open Science Policy

Aligned with the USENIX Security Open Science Policy and our dedication to enhancing the reproducibility and replicability of scientific findings, we commit to sharing all research artifacts related to this paper. These artifacts, including data and code, are available at **https://zenodo.org/records/14719074**. The repository contains the following data and code:

Data. We will make available various types of point cloud data utilized in our research. These data encompass both physical and simulation environments, as detailed below:

- Measurement point clouds (pcd) for each Shadow Material at various distances from the LiDAR.
- Measurement point clouds (pcd) from all experiments.

Code. To support the replication and further development of our research, we will provide the code used in our simulations and for replicating the Shadow Hack. The code includes components such as scripts for simulation setup, attack evaluation, and defense evaluation, as outlined below:

- Optimization code for Shadow Material shapes running on OpenPCDet [30] (Python).
- Code for reproducing Shadow Material (materials not detectable by LiDAR) in AWSIM [20] (C#), ROS2 node (Python).
- Code for BBValidator running on OpenPCDet (Python).

Ethical Considerations

In this study, we proposed an attack method that exploits the physical mechanisms inherent in LiDAR systems and examined its impacts and potential defense strategies. To ensure ethical research conduct, we have taken the following measures.

Collaboration with Affected Stakeholders. We recognized that our research outcomes could affect automobile manufacturers, autonomous driving system developers, and creators of LiDAR datasets. We engaged in collaboration with these stakeholders by initiating a process of responsible disclosure. We have provided them with comprehensive information regarding the potential vulnerabilities identified and corresponding countermeasures. We plan to compile all the details and make them available on our website upon acceptance of this paper.

Assessment of Risks in Data and Model Usage. This research utilized public datasets (KITTI) and open-source object detection models (PointPillars, PointRCNN, SECOND-IoU). Based on the nature of these data, attacks against object detection systems trained on these datasets can be executed. To address this, we are in the process of summarizing precautions for using public point cloud data and sharing this information with data providers and open-source model developers. We are also working to encourage them to alert users about potential risks.

Positive and Negative Impacts of Proposing Attack Methods. The attack method proposed in this study aims to promote the development of defense strategies and enhance the safety of autonomous driving technology, aligning with the public interest as outlined in the Menlo Report [12]. Recognizing the potential misuse of this method, potentially negatively affecting the safety of autonomous vehicles, we carefully considered the benefits and risks. In accordance with the Menlo Report's emphasis on minimizing potential harms, we determined that the potential positive impacts justify the publication of our findings. To further mitigate risks, we collaborated with stakeholders before the publication of this research and recommended the implementation of defense measures.

Safety Considerations for Experiments and Team Members. We used LiDAR devices certified as Class 1 eye-safe in our experiments to eliminate risks to eyesight. Additionally, all experiments were conducted in a safely managed environment within private property, ensuring the safety of team members and the surrounding area.

Compliance with Research Ethics Policies. Based on the USENIX Security '25 Ethics Guidelines and the principles outlined in the Menlo Report, we evaluated potential negative impacts during the research process and at the time of publishing the results, conducting the research with ethical judgment. While the likelihood of this research causing direct harm at this stage is low, we are considering future risks and are engaged in continuous monitoring and responsible disclosure. We believe that the dissemination of this research will ultimately contribute to the public good by improving the security and reliability of autonomous systems.

References

[1] Apollo: Open source autonomous driving. Baidu Apollo team, https://github.com/ApolloAuto/apollo.

[2] Cruise. https://www.getcruise.com/.

[3] RS-LiDAR-M1 User Guide . RoboSense Technology Co., Ltd, https://www.robosense.ai/en/RS-LiDAR-M1.

[4] OS0 ultra-wide view high-resolusion imaging lidar. Ouster, Inc., https://data.ouster.io/downloads/datasheets/datasheet-rev7-v3p1-os0.pdf.

[5] OS1 mid-range high-resolusion imaging lidar. Ouster, Inc., https://data.ouster.io/downloads/datasheets/datasheet-rev7-v3p1-os1.pdf.

[6] Pandar40 40-channel mechanical LiDAR user manual. Hesai Technology Co., Ltd., https://www.hesaitech.com/wp-content/uploads/Pandar40P_User_Manual_402-en-240610.pdf.

[7] QT128 128-channel mechanical lidar user manual. Hesai Technology Co., Ltd., https://www.hesaitech.com/wp-content/uploads/OT128_User_Manual_001-en-240610.pdf.

[8] Velodyne lidar alpha prime. Ouster, Inc., https://data.ouster.io/downloads/datasheets/velodyne/63-9679_Rev-B_DATASHEET_ALPHA-PRIME_web.pdf.

[9] Velodyne Lidar Puck. Ouster, Inc., https://data.ouster.io/downloads/datasheets/velodyne/Puck%20Datasheets.zip.

[10] Waymo. https://waymo.com/waymo-one.

[11] M. Abdelfattah, K. Yuan, ZJ. Wang, and R. Ward. Towards universal physical attacks on cascaded camera-lidar 3d object detection models. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3592–3596, 2021.

[12] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The Menlo Report. *IEEE Security and Privacy*, 10(2):71–75, March 2012.

[13] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[14] Y. Cao, SH. Bhupathiraju, P. Naghavi, T. Sugawara, ZM. Mao, and S. Rampazzi. You can't see me: Physical removal attacks on lidar-based autonomous vehicles driving frameworks. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[15] Y. Cao, N. Wang, C. Xiao, D. Yang, J. Fang, R. Yang, QA. 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 *Proceedings of the 42nd IEEE Symposium on Security and Privacy*, 2021.

[16] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, QA. Chen, K. Fu, and ZM. 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.

[17] Christy, Maggie, NikiNikatos, Phil Culliton, Vinay Shet, and Vladimir Iglovikov. Lyft 3D Object Detection for Autonomous Vehicles. https://kaggle.com/competitions/3d-object-detection-for-autonomous-vehicles, 2019. Kaggle.

[18] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[19] RS. Hallyburton, Y. Liu, Y. Cao, ZM. Mao, and M. Pajic. Security analysis of camera-lidar fusion against black-box attacks on autonomous vehicles. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

[20] TIER IV. AWSIM - open source simulator for self-driving vehicles. https://github.com/tier4/AWSIM, 2022.

[21] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana,

yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - bug fixes and performance improvements, October 2020.

[22] S. Kato, S. Tokunaga, Y. Maruyama, S. Maeda, M. Hirabayashi, Y. Kitsukawa, A. Monrroy, T. Ando, Y. Fujii, and T. Azumi. Autoware on board: Enabling autonomous vehicles with embedded systems. In *ACM/IEEE 9th ICCPS*, 2018.

[23] Ryunosuke Kobayashi, Kazuki Nomoto, Yuna Tanaka, Go Tsuruoka, and Tatsuya Mori. WIP: Shadow Hack: Adversarial Shadow Attack Against LiDAR Object Detection, 2024.

[24] 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 CVPR*, 2019.

[25] nuScenes Forum. Relationship between occlusion and num_points_in_box. https://forum.nuscenes.org/t/relationship-between-occlusion-and-num-points-in-box/249. Accessed on: January 16, 2025.

[26] T. Sato, Y. Hayakawa, R. Suzuki, Y. Shiiki, K. Yoshioka, and QA. Chen. Lidar spoofing meets the new-gen: Capability improvements, broken assumptions, and new attack strategies. In *Proceedings of the Network and Distributed System Security Symposium*, 2024.

[27] 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 CVPR*, 2019.

[28] H. Shin, D. Kim, Y. Kwon, and Y. Kim. Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications. In *Cryptographic Hardware and Embedded Systems–CHES 2017: 19th International Conference*, 2017.

[29] J. Sun, Y. Cao, QA. Chen, and ZM. Mao. Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *Proceedings of the 29th USENIX Security Symposium*, 2020.

[30] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020.

[31] Y. Yan, Y. Mao, and B. Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

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

[33] S. Zhu, Y. Zhao, K. Chen, B. Wang, H. Ma, and C. Wei. Ae-morpher: Improve physical robustness of adversarial objects against lidar-based detectors via object reconstruction. In *Proceedings of the 33rd USENIX Security Simposium*, 2024.

A Appendix

A.1 Surrouding Environments Experiments

Figure 24 shows the five different scenes used in the experiments conducted in Section 6.2.

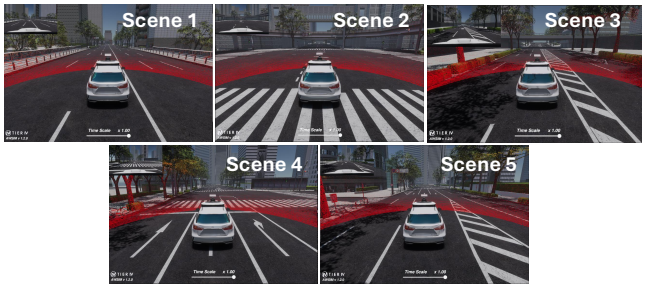

Figure 24: Five distinct driving scenes in AWSIM.

A.2 Shadow Materials with Water Droplets

As mentioned in Section 8.1, environmental conditions such as weather also influence the success of Shadow Hack. For instance, as shown in Figure 25, a mirror sheet with water droplets fails to sufficiently remove point clouds and cannot create an ideal shadow on the point cloud, as seen in Figure 19. Such situations are likely to occur under conditions like rain or dust storms, requiring careful consideration.

A.3 Multi Sensor Fusion

We evaluate Shadow Hack attacks on Multi-Sensor Fusion (MSF), comparing scenarios with and without BBValidator under both attack and non-attack conditions.

Experimental Setup. Using the same setup as in Section 6.1, we positioned a Target Vehicle at a distance of 30 m from the Shadow Material and moved it toward the Shadow Material at a constant speed. We collected point cloud data from 80 consecutive frames and the corresponding onboard camera

Table 3: Detection Performance of MSF Model under Attacked and Non-Attacked Scenarios

MSF Model			Number of Frames with Car Detection		
Images	Point Clouds	Scenario	Images	Point Clouds	MSF
YOLO v5	PointPillars	Attacked	0 / 80	33 / 80	33 / 80
YOLO v5	PointPillars	Non-Attacked	80 / 80	80 / 80	80 / 80
YOLO v5	PointPillars + BBValidator	Attacked	0 / 80	0 / 80	0 / 80
YOLO v5	PointPillars + BBValidator	Non-Attacked	80 / 80	80 / 80	80 / 80

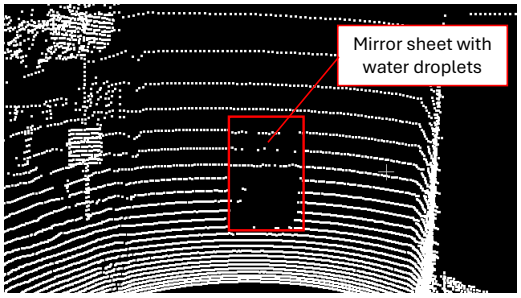

Figure 25: LiDAR scan of mirror sheet with water droplets using OS1-64 sensor.

images until the Target Vehicle reached the Shadow Material, and input them into the MSF model. For comparison, we also collected data in a non-attacked scenario where a vehicle was placed at the location of the Shadow Material in Scene 1. The MSF model consists of PointPillars, a point cloud object detection model trained on the KITTI dataset, and YOLOv5, an image-based object detection model. In this model, an object is detected as a car if either the image or the point cloud detects a car, following an OR-based detection logic. Additionally, we evaluated the case where BBValidator was implemented in PointPillars, setting the BBValidator parameter to $dh = 0.0$ and $N_{\text{thresh}} \geq 19$, which achieved a 100% defense rate as described in Section 7.1.

Results. Table 3 shows the number of frames in which each model combination detected a car in front under each scenario. In the Attacked Scenario, although there is no actual car present, PointPillars erroneously detects a car, as shown in Figure 26, resulting in MSF detecting a car in 33 frames. However, when BBValidator is implemented, the false detections are eliminated, as shown in Figure 27, and consequently, MSF does not detect a non-existent car. In the Non-attacked Scenario, MSF successfully detects the existing car in all frames, regardless of whether BBValidator is implemented. Figure 28 shows an example of the results when BBValidator is implemented. These results demonstrate that BBValidator does not cause false negative detections. Furthermore, even if a false negative detection were to occur due to BBValidator, the car would still be correctly detected from the image, allowing MSF to successfully detect the car overall.

Figure 26: Detection results in the attacked scenario. PointPillars detects a non-existent car.

Figure 27: Detection results in the attacked scenario. PointPillars + BBValidator does not detect the non-existent car.

Figure 28: Detection results in the non-attacked scenario. Both YOLOv5 and PointPillars + BBValidator successfully detect the vehicle in front.