Poster: Shadow Hack: Adversarial Shadow Attack Against LiDAR Object Detection*

Ryunosuke Kobayashi*, Kazuki Nomoto*†, Yuna Tanaka*, Go Tsuruoka*, Tatsuya Mori*‡§,
*Waseda University †Deloitte Tohmatsu Cyber LLC ‡RIKEN §NICT

***Abstract*—Object detection is essential for identifying objects' positions and types from sensor data, especially in autonomous driving systems, where it guides vehicle safety. Machine learning-based object detection, however, faces vulnerabilities, like adversarial samples. This study introduces "Shadow Hack," a novel LiDAR object detection attack method. Unlike previous attacks, which perturb LiDAR data, we strategically place materials like infrared cut films to generate "Adversarial Shadows" on the LiDAR point cloud. This can mislead LiDAR-based object detection in autonomous vehicles, potentially causing congestion and accidents. We evaluate the attack's success rate through simulations and aim to propose countermeasures to enhance system robustness.**

I. Introduction

In recent years, the risk of LiDAR sensor attacks in autonomous vehicles has gained significant attention. These attacks manipulate sensor readings, posing a threat to object recognition systems based on machine learning models. Particularly concerning is the "LiDAR spoofing attack," injecting malicious signals to deceive sensors into detecting nonexistent or missing objects [1, 2]. These attacks target sensors, data processing, and machine learning models, underscoring the imperative to bolster sensor security and enhance model robustness.

This study proposes a new attack vector for sensing systems using LiDAR, named "Shadow Hack," with the aim of understanding its threats and developing effective countermeasures. The concept of this attack lies in exploiting the "shadows" naturally formed in the point cloud data captured by LiDAR sensors (see Figure 1). LiDAR sensors produce point cloud data indicating the presence of objects, but this data also includes the shadows formed behind the objects. Typically, these shadows are ignored in the output of object detection models, but their presence provides important clues for object detection. The Shadow Hack takes advantage of this property of "shadows" by intentionally creating them to fool object detection systems and cause them to malfunction. For example, by placing "Shadow Materials" such as aluminum leisure mats in the environment, false shadows can be created in the point cloud data captured by LiDAR sensors, causing the object detection models to detect non-existent objects (See Figure 2).

II. Shadow Hack Attack Framework

The Shadow Hack attack framework systematically manipulates point cloud data to deceive autonomous vehicles(See Figure 3). First, the attacker collects point cloud data, called X_{benign}, at a specific location (x_t, y_t) along the target vehicle's expected route, using the same LiDAR sensor for data authenticity. Next, the attacker creates an adversarial shadow using the collected X_{benign} data, fine-tuning its location (d_x, d_y) and orientation (θ) to trick the target vehicle's object recognition system. Finally, the attacker implements the Adversarial Shadow by placing materials like infrared-cut film or aluminum leisure mats at the simulated shadow location, making it undetectable by LiDAR. This causes the target vehicle to falsely detect a non-existent object when it encounters the shadow at (x_t, y_t).

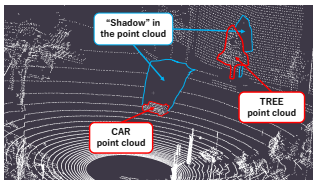


Fig. 1. An example of "Shadow" of the point cloud. "Shadows" are present on the point cloud behind the CAR and TREE.

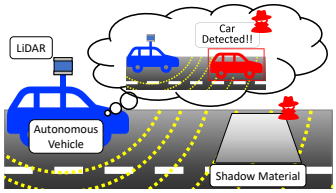


Fig. 2. Attack Overview. Adversarial shadows on the LiDAR point cloud are caused by Shadow Materials like aluminum leisure mats set up by the attacker, resulting in false detection by the autonomous vehicle.

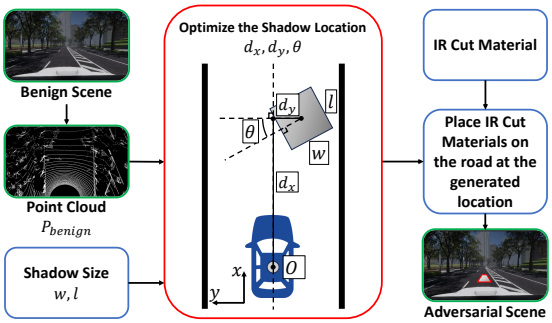


Fig. 3. Shadow Hack Attack Framework

III. Evaluation

We collected point cloud data using AWSIM [3], a simulator designed for autonomous driving software. We used the Ouster OS1-64 LiDAR at 1.73 m height. Data was collected in two environments: a noiseless "Flat" map and urban environments (City 1 - 10). For 3D object detection from point clouds, we employed OpenPCDet [4], evaluating Voxel-based PointPillars and Point-based Point-RCNN models trained on the KITTI dataset. We defined a successful attack as a false detection within 44 m ahead and 3.5 m wide, likely to trigger emer-

*The extended version of this work will be presented at VehicleSec 2024 (WIP).

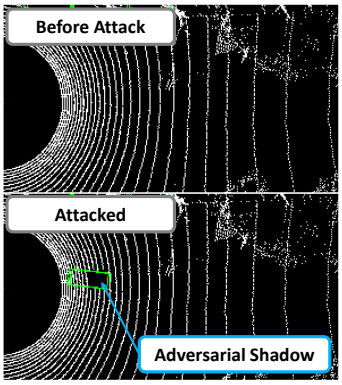


Fig. 4. Object detection results by PointPillars in the City 1 scene

TABLE I
SCENE-WISE ATTACK SUCCESS RATES OF SHADOW HACK ON POINTPILLARS

Scene	Success Rate
Flat	1.00
City 1	0.60
City 2	0.44
City 3	0.04
City 4	0.66
City 5	1.00
City 6	0.88
City 7	0.96
City 8	0.04
City 9	0.20
City 10	0.98

gency braking or steering. A successful attack occurs when a misdetection falls within $0 < d_x \leq 44, -1.25 \leq d_y \leq 1.25$, coinciding with the adversarial shadow, as shown in Fig. 3.

Shadow Hack attacks on PointPillars in City scenes achieved an average attack success rate of 58% when the vehicle was stationary, as presented in Table I. Notably, there were instances where the attack succeeded in all scenes for PointPillars. In contrast, PointRCNN consistently demonstrated an attack success rate of 0% across all scenes. Figure 4 highlights a significant observation regarding PointPillars: its tendency to erroneously identify Adversarial Shadows as objects in specific scenes.

These results underline the impact of differences between the target model and the surrounding environment on the attack success rate. We attribute the observed variations in attack success rates among target models to differences in their internal processing methods. The presence of attack shadows significantly influences object detection outcomes in models that perform inference on point clouds containing shadows. In contrast, models that employ shadow removal techniques prior to object detection exhibit a notably lower attack success rate. PointRCNN, with its low attack success rate of 0%, performs object detection after removing ground points from the point cloud data. Conversely, PointPillars, achieving an attack success rate of 58%, conducts object detection on point clouds that include shadows, without the removal of ground points during processing.

IV. DISCUSSION

Future Work In this paper, we assess the Shadow Hack attack's feasibility, exploiting shadows' impact on LiDAR-based object detection. The observed false positives suggest vulnerabilities in autonomous driving systems. However, our experiments focused on stationary LiDAR, and attacks on moving targets require considering changing shadow shapes. Future research should address this challenge. We evaluate standalone LiDAR object detection models in Section III, showing that the Shadow Hack induces false positives in over half of the frames. However, the real impact on autonomous driving remains unexplored. Therefore, assessing autonomous vehicles' responses to Shadow Hack attacks in both simulation and real-world scenarios is crucial. In simulations, we implemented attack materials directly in AWSIM, evaluating whether autonomous vehicles equipped with Autoware would respond to false positives by stopping or evading. In the real world, we replicated the attack using aluminum leisure mats on Autoware-equipped autonomous vehicles. Regarding the Shadow Hack's shadow generation method, further optimization is possible. Future work will explore angle optimization based on high-confidence false positives during position search, as well as variations in shadow shape and quantity for enhanced attack success rates. Comparative evaluations of different shadow optimization methods are planned.

Countermeasure. As a countermeasure to the Shadow Hack attack, there are three approaches. The Multi-Sensor Fusion Object Detection Model combines LiDAR and camera data to reduce False Positives from the Shadow Hack attack. It first identifies potential object regions using images, then conducts point cloud-based object detection. The Point Cloud Missing Data Detection and Automated Recovery Mechanism detects and fills missing points ("shadows") in point cloud data as a preprocessing step. This counters the effects of the Shadow Hack attack by compensating for anomalies, such as Adversarial Shadows. The Object Detection Model with Tracking Integration uses tracking instead of per-frame detection. While our study focused on single-distance attacks, a tracking-based model may reduce false detections at specific distances. Further investigation is needed to evaluate Tracking Object Detection as a countermeasure.

V. CONCLUSION

This research introduces a novel attack technique called "Shadow Hack," which focuses on manipulating LiDAR point cloud data used in autonomous driving systems to generate artificial "shadows," leading to false object detections. The study evaluates the feasibility of this attack and proposes countermeasures. In simulation tests, Shadow Hack achieved a 58% success rate in causing false detections when applied to the PointPillars object detection model while the vehicle was stationary. Additionally, the study proposes appropriate countermeasures to address the distinctive and unnatural "shadows" generated by Shadow Hack.

ACKNOWLEDGMENT

A part of this work was supported by JSPS KAKENHI 22S0604 and JST CREST JPMJCR23M4.

REFERENCES

[1] Y. Cao et al. "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.

[2] Y. Cao et al. "You Can't See Me: Physical Removal Attacks on LiDAR-based Autonomous Vehicles Driving Frameworks". In: *32nd USENIX Security Symposium*. 2023, pp. 2993–3010.

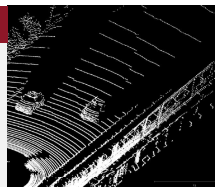
[3] tier4. *AWSIM - Open source simulator for self-driving vehicles*. https://github.com/tier4/AWSIM. 2022.

[4] OpenMMLab. *OpenPCDet: An Open-source Toolbox for 3D Object Detection from Point Clouds*. https://github.com/open-mmlab/OpenPCDet. 2020.

Poster: Shadow Hack: Adversarial Shadow Attack Against 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

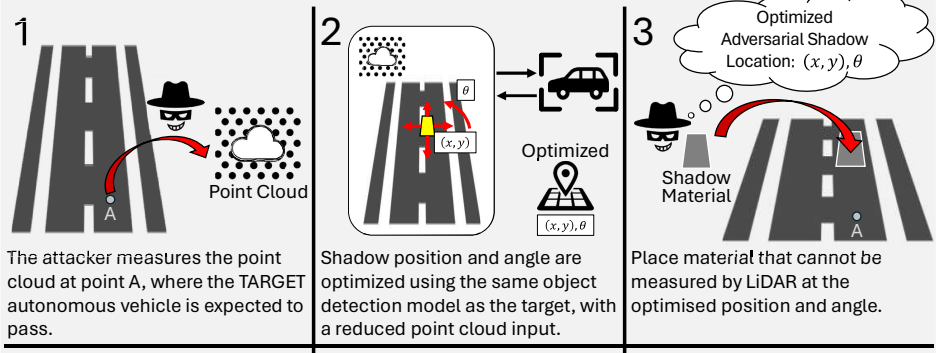
Introduction



- The point cloud object detection model is trained on a dataset that contains "**shadows**."
- While the proposed adversarial attacks primarily focus on adding perturbations to the point cloud, there is a possibility that "shadows" also impact object detection.
- We propose a "**Shadow Hack**" attack that induces false detections by replicating precomputed "shadows" on the point cloud.

Overview of the Shadow Hack Attack Framework

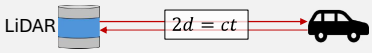
Key idea: By placing materials that LiDAR cannot measure on the ground, create an Adversarial Shadow.



1. The attacker measures the point cloud at point A, where the TARGET autonomous vehicle is expected to pass.
2. Shadow position and angle are optimized using the same object detection model as the target, with a reduced point cloud input.
3. Place material that cannot be measured by LiDAR at the optimised position and angle.
4. When the target passes through A, an Adversarial Shadow appears in the target's measurement point cloud.
5. The TARGET object detection model causes false positives due to Adversarial Shadow.

LiDAR（Light Detection And Ranging）

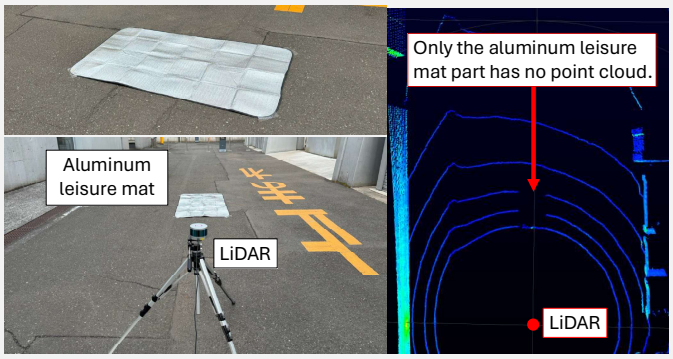
- Time of Flight: Distance is measured using reflected infrared laser beams.

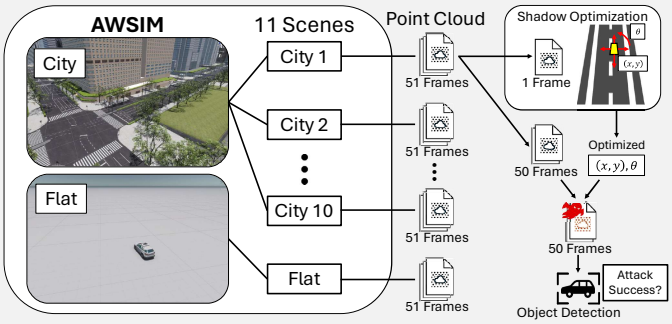


t: time[s]
d: distance[m]
c: speed of light[m/s]

Shadow Materials that LiDAR cannot measure

- Reflection (Mirror, Aluminum leisure mat, IR Cut filter)
- Absorption (Infrared absorbing cloth)
- Transparency (Glass)





Experiment Setup

Simulator: AWSIM (Designed for developing Autoware) **LiDAR:** Ouster OS1-64
Target Model: PointPillars, PointRCNN (Pretrained using the KITTI dataset)

Optimization of shadows is performed in one frame out of 51 frames, and the remaining 50 frames are used for replication in each of the 11 scenes: Flat, City1 – 10, to evaluate the attack success rate.

Results

- Against **PointPillars**, the attack success rate was 100% in Flat and the average attack success rate in City was 58%.
- Against **PointRCNN**, the attack success rate was 0% in all scenarios.

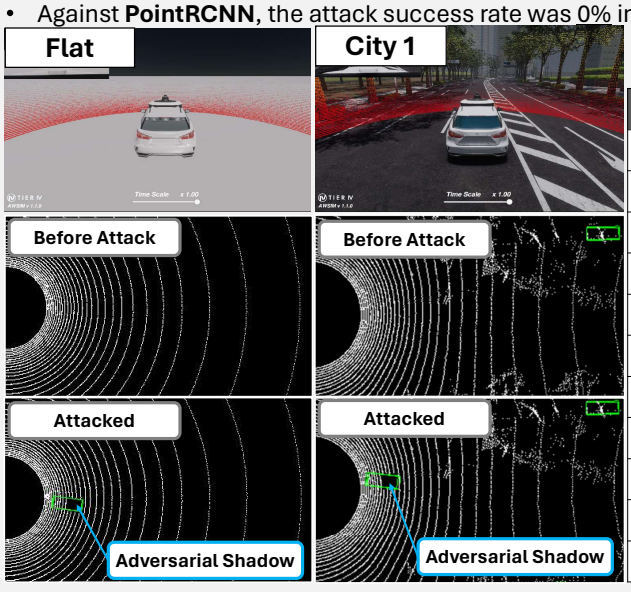
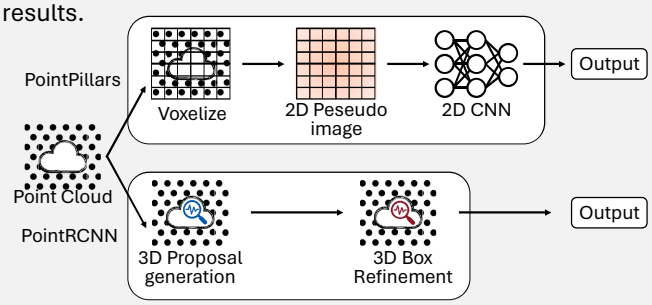


Table: Attack Rates on PointPillars

Scene	Success Rate
Flat	1.00
City 1	0.60
City 2	0.44
City 3	0.04
City 4	0.66
City 5	1.00
City 6	0.88
City 7	0.96
City 8	0.04
City 9	0.20
City 10	0.98

Discussion of the Results

Object detection architecture could be critical to the attack results.



Countermeasure

1. Multi-Sensor Fusion Object Detection Model
2. Point Cloud Missing Data Detection and Automated Recovery Mechanism
3. Object Detection Model with Tracking Integration

Future Work

- Attack Capabilities on Moving Autonomous Vehicles
- Impacts on Autonomous Driving Systems
- Extension of the Shadow Optimization Process