

USENIX

THE ADVANCED COMPUTING
SYSTEMS ASSOCIATION

# Seeing is Deceiving: Systematic Vulnerability Analysis of LiDAR-Based Autonomous Driving to Mirror-Induced Perception Failures

Selma Yahia and Ildi Alla, *Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg;* Girija B. Mohan and Daniel Rau, *University of Applied Science Saarland (htw saar);* Mridula Singh, *CISPA Helmholtz Center for Information Security;* Valeria Loscri, *Inria Lille - Nord Europe*

https://www.usenix.org/conference/vehiclesec26/presentation/yahia

This paper is included in the Proceedings of the
4th USENIX Symposium on Vehicle Security and Privacy.

August 10–11, 2026 • Baltimore, MD, USA

ISBN 978-1-939133-56-4

Open access to the Proceedings of the
4th USENIX Symposium on Vehicle Security and Privacy
is sponsored by USENIX.

# Seeing is Deceiving: Systematic Vulnerability Analysis of LiDAR-Based Autonomous Driving to Mirror-Induced Perception Failures

Selma Yahia* Ildi Alla* Girija B. Mohan† Daniel Rau† Mridula Singh‡ Valeria Loscri§

*Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg
†University of Applied Sciences Saarland (htw saar), 66117 Saarbrücken, Germany
‡CISPA Helmholtz Center for Information Security
§Inria Lille – Nord Europe

## Abstract

Autonomous vehicles rely heavily on LiDAR for accurate 3D perception; however, this sensing modality remains vulnerable to several physical-world failure modes. This paper presents a systematic vulnerability analysis of LiDAR-based autonomous driving stacks against mirror-induced perception failures. We identify, formalize and evaluate two distinct failure modes arising from planar mirrors in the environment: *Object Addition*, where multi-path reflections create phantom obstacles in free space, and *Object Removal*, where beam deflection renders real obstacles invisible to the sensor. Our analysis makes three contributions. First, we conduct an end-to-end vulnerability assessment on a full-stack autonomous vehicle running Autoware, tracing how mirror-induced artifacts propagate from raw perception through planning and control, ultimately triggering emergency braking and cascade collisions. Second, we derive physics-informed empirical models that characterize artifact location, point count, and appearance probability as functions of distance, angle, and mirror area. Third, we integrate these models into a CARLA-based simulation framework for real-time artifact injection, addressing a key limitation of current simulators that do not model specular reflections. Our findings reveal that even small mirrors ($0.18\,m^2$, costing under \$60) can cause high-confidence false detections (74% "Vehicle" classification) and trigger unsafe driving behaviors. We discuss potential countermeasures and highlight that robust defenses against physically legitimate but geometrically deceptive sensor inputs remain an open challenge.

## 1 Introduction

The deployment of autonomous driving systems on public roads depends on the ability of their perception pipelines to construct an accurate, trustworthy model of the physical world. Among the sensor modalities employed, Light Detection and Ranging (LiDAR) plays a central role, providing high-resolution 3D point clouds that form the basis for obstacle detection, localization, and path planning [14, 30]. The integrity of LiDAR data is therefore not merely a performance concern but a safety-critical requirement. Errors in the raw sensor data can propagate through the entire autonomy stack, from perception through planning to control, with potentially catastrophic consequences.

A substantial body of research has investigated LiDAR vulnerabilities through *active* spoofing attacks, which use external laser emitters to inject forged pulses into the sensor [10, 18, 19, 21, 26], or through *passive* adversarial objects, such as 3D-printed shapes or specially crafted road patterns designed to exploit weaknesses in learned perception models [1, 24, 27]. Active attacks require specialized hardware and precise synchronization with the victim sensor. Passive adversarial objects typically target specific neural network architectures rather than the fundamental sensing modality.

However, a more fundamental class of vulnerability has received remarkably little attention: the susceptibility of LiDAR to deception through ordinary specular reflection. The robotics community has long recognized that mirrors and reflective surfaces create problematic artifacts in LiDAR data [3, 28], and recent work has begun to explore reflective materials for targeted point removal [11, 12]. Yet the broader implications of this physical phenomenon, particularly the ability of simple mirrors to both *create* phantom objects and *erase* real ones, have not been systematically analyzed in the context of full-stack autonomous driving systems.

This vulnerability is particularly concerning for several reasons. First, unlike active spoofing, mirror-based artifacts are generated from the vehicle's own laser pulses, making the returned signals *physically legitimate* in terms of wavelength and temporal structure. Defenses designed to reject externally injected pulses are therefore ineffective against such physically consistent reflections. Second, the artifacts are structured and spatially coherent, following the deterministic laws of geometric optics rather than appearing as random noise. Third, mirror placement requires no electronics, no precise synchronization, and no knowledge of the victim system's internal parameters. Only general knowledge of road geometry and

Table 1: Comparison of mirror-based LiDAR vulnerability analysis with prior work on LiDAR deception.

| Work | Mirror-Based | Passive | Object Addition | Object Removal | Geometric Modeling | Simulation | Real-World Validation | AV Stack Evaluation | Defense Proposal |
|---|---|---|---|---|---|---|---|---|---|
| **Ours** | ● | ● | ● | ● | ● | CARLA | Outdoor (Vehicle) | Autoware | ● |
| [24] | ○ | ● | ● | ○ | ○ | ○ | Indoor (Basic) | ○ | ○ |
| [1] | ○ | ● | ● | ○ | ○ | ○ | Outdoor | ○ | ○ |
| [27] | ○ | ● | ● | ○ | ● | LGSVL | Outdoor | ○ | ● |
| [11] | ● | ● | ○ | ● | ○ | AWSIM | ○ | ○ | ● |
| [12] | ● | ● | ○ | ● | ○ | AWSIM | Outdoor (Basic) | ○ | ● |
| [31] | ○ | ● | ○ | ○ | ● | SVL | Outdoor | ○ | ● |
| [2] | ○ | ○ | ● | ○ | ○ | Baidu Apollo | Indoor (Basic) | ○ | ● |
| [22] | ○ | ○ | ● | ○ | ● | Dataset | Indoor (Basic) | ○ | ● |
| [32] | ○ | ● | ○ | ● | ● | Dataset | Outdoor (Drones) | ○ | ● |

typical sensor mounting configurations is required.

Table 1 provides a comparative overview of prior work. To our knowledge, prior research has not provided a systematic, end-to-end analysis of mirror-induced LiDAR deception in autonomous driving systems. We identify four specific gaps: **(G1)** Prior studies evaluated reflective artifacts only at the point-cloud or perception-model level [2, 22, 31], without tracing their propagation through a complete autonomy pipeline to quantify system-level safety implications. **(G2)** No empirical or analytical models exist to describe how artifact properties depend on controllable physical parameters such as angle, distance, and area, preventing systematic prediction or reproduction [3, 28]. **(G3)** Current LiDAR simulators [4] rely on simplified single-bounce ray-casting that does not model specular multi-path reflections. Consequently, mirror-induced artifacts cannot be realistically reproduced in simulation, limiting controlled and repeatable evaluation. **(G4)** Recent mirror-related work [11, 12] focused narrowly on point removal, without exploring the dual capability of mirrors to also *inject* structured phantom objects.

**This Work.** We present the first end-to-end vulnerability analysis of LiDAR-based autonomous driving systems against mirror-induced perception failures. Our approach follows a structured methodology consisting of real-world vulnerability discovery, physics-informed modeling, simulation-based scalable evaluation, and full-stack impact validation. Specifically, we make the following contributions:

- **End-to-End Vulnerability Assessment (G1):** We evaluate the impact of mirror-induced artifacts on a fully integrated autonomous vehicle running the Autoware stack with a commercial LiDAR. We trace how artifacts propagate through perception, from point cloud to occupancy grid to object detection, and into planning and control. We demonstrate that these artifacts trigger false emergency braking, high-confidence phantom vehicle classifications of up to 74%, and cascade collision scenarios.
- **Physics-Informed Empirical Models (G2):** We derive and validate four analytical models that characterize mirror-induced LiDAR artifacts, including radial distance, lateral offset, point count, and appearance probability, as functions of distance, mirror angle, and area. These models are grounded in geometric optics and calibrated through controlled outdoor experiments.
- **Simulation Framework for Specular Artifacts (G3):** We implement a real-time, model-driven artifact injection pipeline in CARLA that generates physically realistic mirror artifacts during simulation. This framework bridges the sim-to-real gap for specular reflections and enables safe, repeatable testing of diverse configurations.
- **Comprehensive Threat Characterization (G4):** We formalize two complementary failure modes, Object Addition, defined as phantom obstacle injection, and Object Removal, defined as real obstacle concealment. We demonstrate these failure modes through a combination of real-world experiments and simulation, providing a comprehensive characterization of the vulnerability surface, as illustrated in the end-to-end pipeline shown in Figure 1.

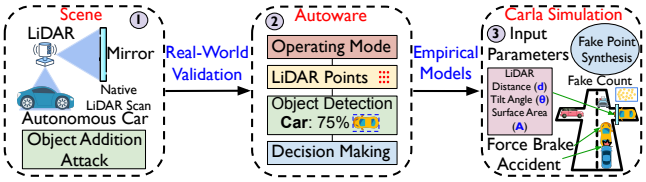



Figure 1: End-to-end evaluation pipeline. (①) Real-world experiments characterize mirror-induced artifacts and produce empirical models. (②) Key outcomes are validated on an Autoware-equipped vehicle, confirming perception failures and unsafe driving behaviors. (③) Empirical models drive real-time artifact injection in CARLA for scalable evaluation.

**Responsible Disclosure.** All experiments were conducted in a controlled environment with no risk to public safety. We disclose this vulnerability to support the development of reflection-aware perception and validation systems. Demon-




Figure 2: Types of reflection in LiDAR sensing. (a) Diffuse: incident rays scatter broadly on rough surfaces. (b) Specular: a smooth mirror reflects coherently with $\theta_r = \theta_i$.

stration videos and supplementary materials are available at: https://lidar-attack.github.io/lidar-research/.

## 2 Background

### 2.1 LiDAR Operation and Reflection Physics

LiDAR sensors emit pulses of near-infrared laser light and compute the distance to a surface using the time-of-flight (ToF) principle: $d_l = \frac{c \cdot \Delta t}{2}$, where $c$ is the speed of light and $\Delta t$ is the measured round-trip time. Contemporary systems scan the environment using rotating or solid-state arrays with fixed angular resolution, producing a 3D point cloud $P = \{(x_i, y_i, z_i, r_i)\}_{i=1}^{N}$, where $(x_i, y_i, z_i)$ is the 3D location and $r_i$ the return intensity. Point cloud fidelity is therefore tightly coupled to how beams interact with physical surfaces.

Two surface interaction modes determine LiDAR returns (Figure 2). *Diffuse reflection* occurs on rough surfaces, scattering return energy broadly and allowing detection from multiple angles. *Specular reflection* occurs on smooth, mirror-like surfaces (glass, wet roads, glossy vehicle bodies) and follows the law of reflection:

$$\vec{v}_{\text{ref}} = \vec{v}_{\text{in}} - 2\,(\vec{v}_{\text{in}} \cdot \vec{n})\,\vec{n}, \tag{1}$$

where $\vec{v}_{\text{in}}$ is the incoming ray and $\vec{n}$ is the surface normal. This specular behavior is the source of the measurement artifacts central to this work. We note that real mirrors absorb a small fraction of incident energy at each reflection (typically 3–8% per bounce for silver-backed glass). For the two-hop paths in our scenarios, the cumulative energy loss remains below the sensor's detection threshold in most configurations, as validated by our experimental data. At greater path lengths or with additional bounces, absorption could reduce artifact intensity below detectable levels, a factor captured by the distance-dependent attenuation in our empirical models.

### 2.2 Mirror-Induced LiDAR Measurement Artifacts

When a LiDAR beam strikes a specular surface, its path is redirected according to the law of reflection. This can corrupt the resulting point cloud through two primary mechanisms:
**Data Omission** (Figure 3a) occurs when a mirror deflects a LiDAR beam away from the sensor, preventing a return signal.

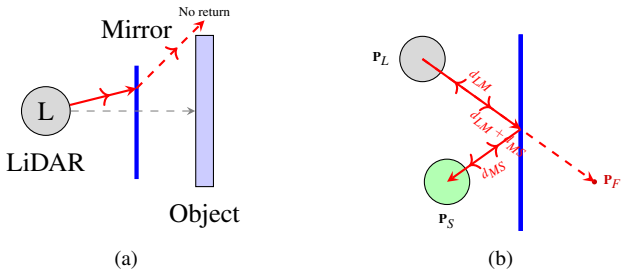



Figure 3: Primary LiDAR measurement artifacts from specular reflections. (a) Data omission: beam reflected away, no return. (b) Data fabrication: two-hop path creates a virtual point at $\mathbf{P}_F$.

A beam that would normally strike a real object ($\mathbf{P}_O$) instead reflects along a path that never returns, causing the LiDAR to record empty space. The object is effectively removed from the point cloud.
**Data Fabrication** (Figure 3b) occurs when a LiDAR beam is redirected by a mirror, strikes a secondary surface ($\mathbf{P}_S$), and returns to the sensor via the same mirror. The LiDAR assumes a direct line of sight and misinterprets the total travel time:

$$d_{\text{measured}} = \frac{d_{\text{total}}}{2} = d_{LM} + d_{MS}, \tag{2}$$

where $d_{LM}$ is the LiDAR-to-mirror distance and $d_{MS}$ is the mirror-to-secondary-surface distance. The system projects this total distance along the original emission direction, placing the reflected object at a false location $\mathbf{P}_F$, corresponding to the geometric mirror image of the actual surface. This creates phantom obstacles in empty space.

## 3 Threat Model

Urban driving environments contain numerous specular surfaces, including glass storefronts, building facades, bus shelters, and traffic convex mirrors, that can redirect LiDAR pulses along unintended paths. While such reflections occur naturally and may cause transient artifacts, the underlying vulnerability can also be *deliberately exploited* by an adversary who controls the placement or orientation of a reflective surface. This section formalizes the threat across a spectrum from incidental to adversarial.

### 3.1 Adversary Goals

The adversary seeks to introduce physically plausible but semantically deceptive artifacts into the AV's LiDAR point cloud, leading to unsafe or disruptive driving behavior. We define two complementary failure modes:
• **Object Addition (OA):** Create a *false positive* by fabricating a phantom obstacle in the AV's path, aiming to trigger an inappropriate defensive response such as emergency braking in flowing traffic or an unnecessary evasive maneuver.

• **Object Removal (OR):** Create a *false negative* by making a real obstacle invisible to the LiDAR sensor, causing the AV to fail to detect a genuine hazard.

Both failure modes produce structured, plausible point clouds rather than obvious noise, making them difficult to distinguish from legitimate sensor data at the perception level.

## 3.2 Adversary Capabilities and Constraints

The adversary operates as a *passive, external* agent in the physical domain with the following profile: (1) **Black-box access:** no internal access to the AV's software, sensor data, or control systems; the manipulation is non-invasive, relying only on observable physical reactions. (2) **Passive mechanism:** no electronic emitters are used; unlike active spoofing [20, 21], the adversary redirects the victim's own LiDAR pulses, producing physically legitimate but geometrically deceptive returns. (3) **General environmental knowledge:** the adversary knows the road layout and typical LiDAR mounting positions but not the specific vehicle's calibration parameters. (4) **Physical mirror placement:** one or more portable planar mirrors are placed before the AV, potentially disguised as infrastructure elements (utility boxes, construction signs, roadside art).

**Adversary Controllability.** The adversary's ability to manipulate phantom artifacts is constrained by the underlying two-hop reflection geometry (Section 2), yet remains sufficiently structured to enable practical attacks. According to the law of reflection (Equation 1), the adversary can act on three physical variables: mirror placement, mirror orientation (tilt angle $\theta$), and reflective area $A$. Specifically, *phantom location* is governed by the combination of mirror placement and orientation. Mirror placement sets the LiDAR-to-mirror distance $d$, which in turn scales the lateral displacement $d\tan(2\theta)$ and the apparent depth along the two-hop path (Equation 2). Consequently, by positioning the mirror at a chosen standoff distance from the victim's trajectory and selecting an appropriate tilt, the adversary can place the phantom within the victim's planned path (e.g., at an intersection turning corridor). *Phantom density*, quantified as the number of reflected points per frame, is primarily controlled by the mirror area $A$. Larger surfaces intercept a greater fraction of the divergent LiDAR beam and produce denser clusters.

*Phantom shape* is not directly chosen by the adversary but is determined by the mirror image of whatever secondary surface the deflected beam strikes. The secondary surface is therefore environmental rather than adversary-controlled. In typical urban deployments, walls, facades, parked vehicles, and the ground plane provide abundant candidates at predictable distances from intersection corners. Even in the absence of such infrastructure, the ego vehicle itself can serve as the secondary surface, yielding a naturally vehicle-shaped phantom that might be misclassified by the perception stack. Consequently, the feasibility of the attack does not hinge on fine-grained environmental control by the adversary.

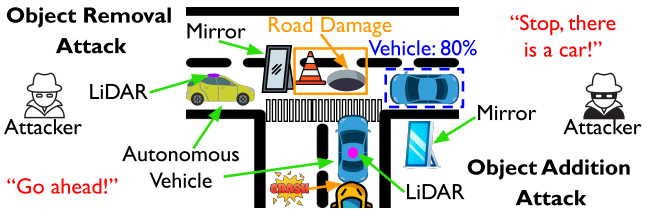



Figure 4: Illustrative scenarios. **OR**: A mirror redirects LiDAR beams away from a traffic cone, creating the illusion of a clear path. **OA**: A mirror reflects pulses to create a phantom obstacle, triggering a false positive.

## 3.3 Threat Spectrum: Incidental to Adversarial

Mirror-induced LiDAR artifacts can arise across a spectrum of intent, from incidental environmental reflections to deliberate adversarial placement of reflective surfaces.

• **Non-adversarial (environmental):** Glass building facades, store windows, wet road surfaces, and metallic infrastructure can produce unintended specular returns.

• **Semi-adversarial (repurposed infrastructure):** Convex mirrors are routinely deployed at blind curves and narrow intersections for human driver visibility. An adversary could subtly re-angle such a mirror/install an additional one that appears functionally legitimate. Because convex mirrors are *expected* in these environments, no modifications are required.

• **Fully adversarial (dedicated placement):** Purpose-built reflector deployment, intended to induce dense and spatially coherent artifacts, mounted on roadside infrastructure or disguised as a traffic convex mirror at an intersection corner.

## 3.4 Representative Scenarios

We ground the threat model with two deployment scenarios, illustrated in Figure 4:

• **Scenario 1: Intersection Blocking via Object Addition.** A mirror is positioned at an urban intersection corner, oriented toward the approaching AV's expected trajectory. This mirror causes the LiDAR sensor to register phantom points in the AV's turning path, where no real obstacle exists. The perception system interprets these as a legitimate obstruction, triggering unnecessary braking or evasive maneuvers.

• **Scenario 2: Obstacle Concealment via Object Removal.** A mirror is placed at the roadside to intercept LiDAR beams that would otherwise detect a real obstacle (e.g., a construction barrier). Instead of returning from the obstacle, the sensor receives returns consistent with a clear road surface. The AV consequently infers a traversable path through an actually obstructed region.

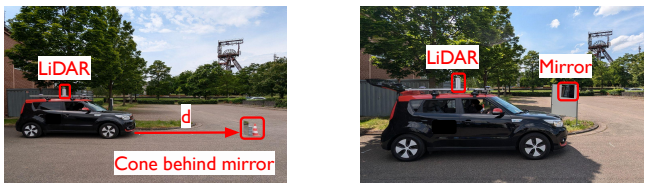



Figure 5: Experimental setups. **(a)** Object Removal: a mirror obscures a traffic cone from the LiDAR's line of sight. **(b)** Object Addition: a mirror array redirects LiDAR pulses toward a nearby structure, creating phantom returns in the vehicle's path.

## 4 Real-World Experimental Campaign

This section describes the controlled outdoor experiments we conducted to characterize the physical behavior of mirror-induced LiDAR artifacts. The experiments serve two purposes: (1) providing empirical evidence of the vulnerability on a real sensor platform, and (2) generating the dataset from which our empirical models (Section 6) are derived.

### 4.1 Platform and Setup

All experiments used a test vehicle with a roof-mounted *Ouster OS1-128* LiDAR sensor at approximately 2.2 m height, operating at 10 Hz with 128 vertical channels and full 360° horizontal coverage [16]. Data was recorded using ROS 2, logging raw LiDAR packets for precise offline analysis.

We used optically flat, silver-backed glass mirrors mounted on adjustable stands with angular protractors for tilt control. Two mirror configurations were employed: single panels of approximately 60 cm×40 cm for OR tests, and modular arrays of 30 cm×30 cm tiles (2, 4, or 6 tiles, yielding 0.18, 0.36, and 0.6 $m^2$ of reflective area) for OA tests. Total cost of all mirror hardware was approximately $60.

Testing was conducted in a closed campus parking lot and adjacent roadway under clear, dry daytime conditions. The vehicle was driven at speeds up to 8 km/h to replicate low-speed approach scenarios (e.g., intersection turns, parking), enabling capture of the full temporal evolution of artifacts across multiple LiDAR frames as the geometry between vehicle and mirror changed.

### 4.2 Object Removal Trials

In the OR setup (Figure 5a), a traffic cone (50 cm tall, 29 cm base) was placed 4 m from the vehicle. A single mirror (60×40 cm) was positioned between the LiDAR and the cone, covering the cone's vertical extent as seen by the sensor. We tested mirror tilt angles of 0°, 15°, 30°, and 45° relative to the LiDAR's forward axis. Each configuration was recorded for 10 seconds (∼100 LiDAR frames), with the vehicle stationary to isolate the mirror's effect.

**Observation.** Across all tested configurations, the cone was completely absent from the point cloud when the mirror was present (Figure 6a and Figure 6b). This demonstrates that a mid-sized obstacle can be rendered fully invisible through passive beam redirection. Critically, this failure mode is fundamentally different from simple occlusion by an opaque object or from beam redirection into empty space: the mirror deflects pulses toward the ground plane, where they produce valid-looking ground returns that are spatially continuous with the surrounding floor (Figure 6c). Rather than leaving a visible data void, the obstructed region is populated by geometrically consistent returns, causing the sensor to perceive a continuous traversable surface where the cone is located. This distinction is what makes OR particularly insidious. Void-based anomaly detectors that reason about missing returns or shadow regions [8] can catch opaque occlusion, but cannot catch this substitute-surface mechanism where the voided region is replaced by geometrically consistent ground returns.

### 4.3 Object Addition Trials

In the OA setup (Figure 5b), a modular mirror array was mounted on a vertical board at an intersection edge. We tested three tilt angles (30°, 45°, 60°) and three reflective areas (0.18, 0.36, 0.6 $m^2$), yielding a 3×3 parameter grid. The vehicle approached from ∼20 m at 8 km/h, executing an intersection turn while continuously recording LiDAR data.

**Observation.** Virtual points behind the mirror were consistently observed across all trials (Figure 7). Two key findings emerged. First, **mirror area controls artifact density**: increasing from 0.18 $m^2$ to 0.6 $m^2$ transformed the artifact from a sparse cluster into a dense, cohesive structure resembling a solid object. Second, **mirror angle controls artifact timing and location**: shallow angles (30°) induced abrupt, close-range phantoms that emerged at short LiDAR-to-mirror distances, whereas steep angles (60°) produced more spatially persistent artifacts detectable earlier in the vehicle's approach and at larger perceived depths, as shown in Figure 8 (extended analysis is provided in Appendix B.1). The 45° case exhibited intermediate behavior with noticeable frame-to-frame variability in artifact persistence and spatial consistency. These observations reveal a key trade-off: *angle* determines artifact *placement*, while *area* determines artifact *density* and *perceptual plausibility*.

## 5 Full-Stack Vulnerability Assessment on Autoware

This section presents the core vulnerability assessment: tracing how mirror-induced artifacts propagate through a fully integrated autonomous driving system and trigger safety-critical failures. Building on the experimental evidence from the previous section, we demonstrate the end-to-end impact on perception, planning, and control using an Autoware-equipped

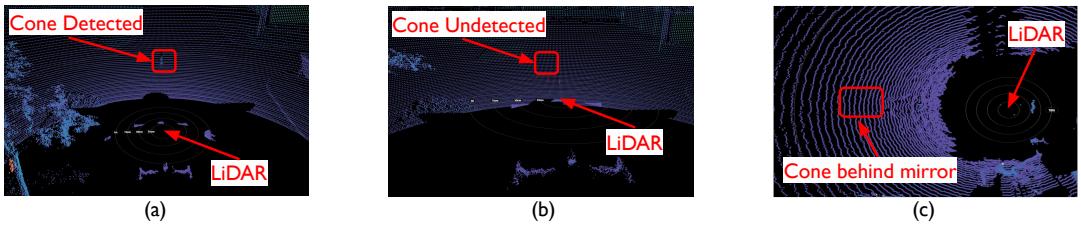



Figure 6: Object Removal at 0° mirror tilt. **(a)** Baseline: cone visible. **(b)** Side view with mirror: cone completely removed from perception. **(c)** Top view with mirror: the region where the cone was located is filled with ground returns that are spatially consistent with the surrounding floor, rather than showing a data void.

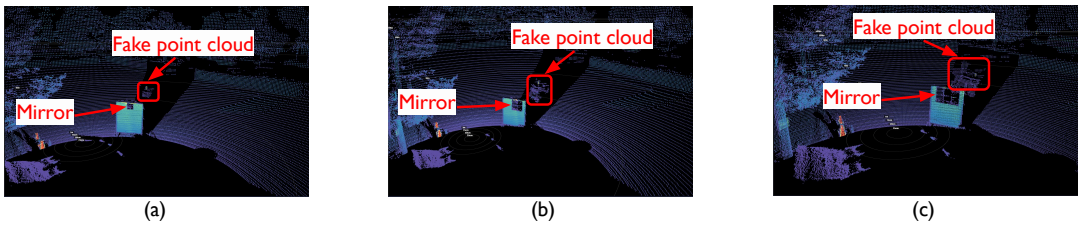



Figure 7: OA point clouds with varying mirror area. **(a)** 2 tiles: sparse cluster (circled). **(b)** 4 tiles: denser vertical surface. **(c)** 6 tiles: high-density cluster approximating a vehicle profile.

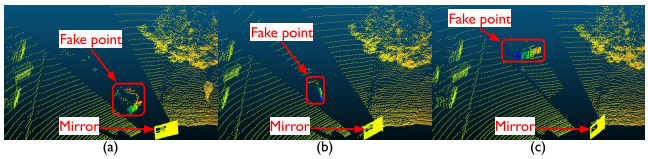



Figure 8: Impact of mirror tilt angle on phantom location: **(a)** 30°; **(b)** 45°; **(c)** 60°.

test vehicle, showing that phantom points at the sensor level cascade into emergency braking, unnecessary stops, and disrupted trajectory planning.

## 5.1 Methodology

The validation used the same Kia Soul EV with Ouster OS1-128 LiDAR described in Section 4, running the Autoware.Universe stack with CenterPoint [29] as the primary 3D object detector. CenterPoint was selected for its strong performance on standard benchmarks and its deployment in real-world AV systems. Experiments were conducted in two modes:

- **Manual Mode:** A safety driver maintained precise trajectories while logging perception outputs, enabling clean analysis of how the perception stack interprets adversarial inputs.
- **Autonomous Mode:** The full autonomy stack (perception, planning, control) was engaged, allowing observation of the system's emergent behaviors in response to artifacts.

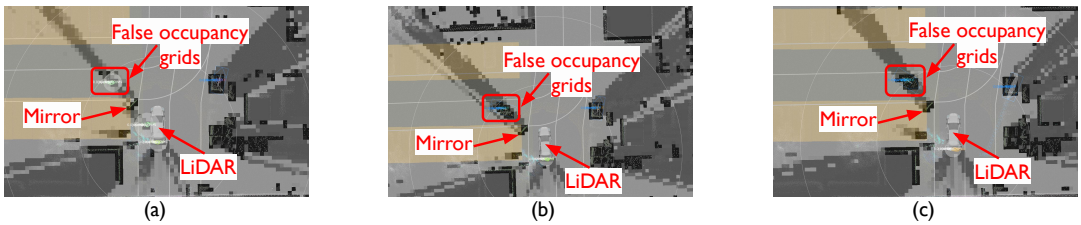
Each scenario was tested across the full parameter grid (three mirror areas × three tilt angles). Synchronized logs captured occupancy grids, detected object lists, and vehicle control signals.

## 5.2 Perception-Level Failure Analysis

① **Object Addition: Phantom Detection and Misclassification.** Under OA conditions, mirror-reflected returns corrupt the perception pipeline at multiple levels. In the occupancy grid (Figure 9), two mirror tiles mark several cells as "occupied." With four and six tiles, the occupied region grows to match a vehicle-sized footprint, persisting across multiple frames as the mapping module accumulates occupancy evidence from consecutive LiDAR scans.

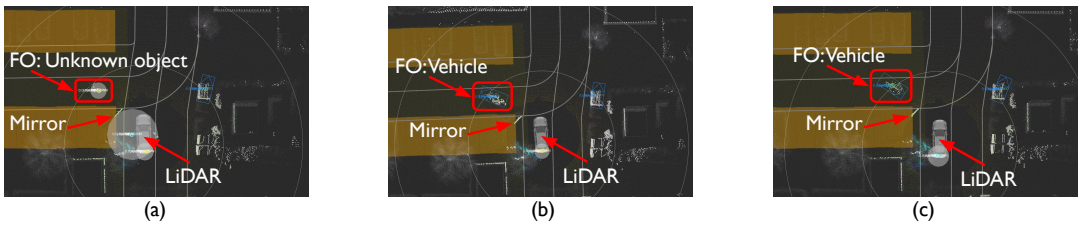
At the semantic level, the corruption escalates from spatial to categorical (Figure 10). With two tiles, the sparse cluster is classified as "UNKNOWN", but still treated as an obstacle by downstream modules. With four tiles, the denser artifact triggers a "CAR" classification at 65% confidence. With six tiles, confidence rises to 74%. The detector progressively hallucinates a plausible, high-confidence vehicle where none exists, revealing a fundamental breakdown in perception integrity: the physically legitimate but geometrically deceptive signals are indistinguishable from real objects within the detector's learned feature space.

② **Object Removal: Silent Obstacle Erasure.** The OR scenario introduces a particularly insidious failure: the mirror redirects LiDAR beams away from the cone and toward the ground plane, where they produce geometrically consistent returns. These reflected signals are processed as structurally valid measurements, and Autoware's ground segmentation interprets them as part of a continuous, traversable surface. As shown in Figure 11, the baseline correctly marks the cone as occupied (black cell), while the mirror condition shows the same region as free space despite the physical presence of both the cone and the mirror. This data replacement makes the OR vulnerability particularly dangerous. It can pass through



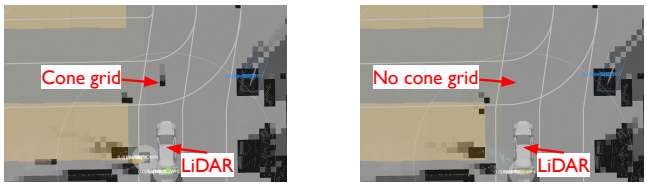

(a) (b) (c)

Figure 9: Occupancy grid corruption under Object Addition. **(a)** 2 tiles: thin line of occupied cells behind mirror. **(b)** 4 tiles: dense cluster triggers "CAR" detection. **(c)** 6 tiles: occupied area matches a real vehicle footprint.




(a) (b) (c)

Figure 10: CenterPoint detection outcomes under Object Addition. **(a)** 2 tiles: "UNKNOWN" detection (white circle). **(b)** 4 tiles: "Vehicle" at 65% confidence (blue box). **(c)** 6 tiles: "Vehicle" at 74% confidence.




(a) (b)

Figure 11: Occupancy grid under Object Removal. **(a)** Baseline: cone correctly marked occupied. **(b)** With mirror: same region falsely marked free.

anomaly detection systems that search for sensor voids or dropouts, and it compromises safety architectures that treat LiDAR as ground-truth for obstacle detection.

## 5.3 System-Level Behavioral Failures

Having characterized the perception-level impact, we now examine how OA artifacts propagate to planning and control in **autonomous mode**. We focus on OA using the 6-tile configuration (maximum artifact density) across three mirror angles, with the vehicle following a route toward the mirror setup at 8 km/h.

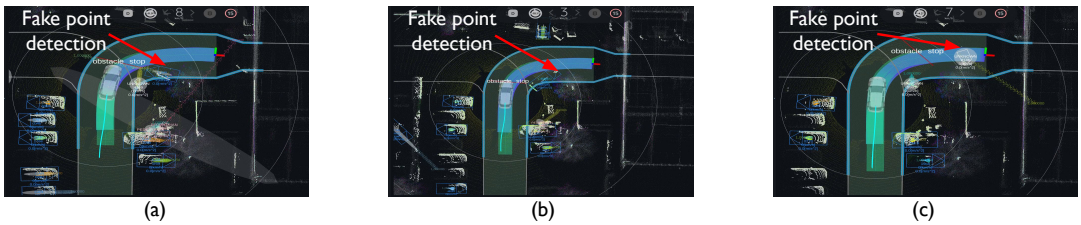
**Observed Failures.** In every trial, the planner interpreted the phantom as a real threat and engaged its *obstacle stop* procedure. However, the nature of the failure varied significantly with mirror angle, as shown in Figure 12, producing a spectrum of dangerous behaviors:

• **Late-Onset, High-Jerk Braking (30°):** At a shallow 30° angle, the reflective geometry only produced a detectable phantom when the vehicle was within a few meters of the mirrors. As shown in Figure 12a, the vehicle has almost completed the turn's apex when the phantom object (blue box) appears abruptly at extremely close range. This late-onset detection created a critically low Time-to-Collision (TTC), forcing the motion planner to issue an immediate emergency *obstacle stop* command. The vehicle halted with an extremely high jerk profile, a maneuver that, while logically correct from the system's perspective, would be highly dangerous in a real traffic environment and would likely cause a rear-end collision by a following vehicle.

• **Mid-Turn Stop with Decision Instability (45°):** Increasing the angle to 45° widens the window for reflection. Consequently, the phantom object is detected earlier in the vehicle's approach, as it begins to enter the curve but before it reaches the apex. The RViz view shows (see Figure 12b) the vehicle braking while still on the initial part of the turn. The planner has slightly more time to react compared to the 30° case, but the outcome is still a full stop in the middle of a dynamic maneuver. This scenario can be particularly hazardous if it leads to detection flickering, causing decision instability (oscillating between braking and proceeding) and creating an unpredictable hazard for other road users.

• **Pre-Turn Path Obstruction (60°):** At 60°, the mirror presents a large, effective reflective surface to the approaching vehicle. This results in the earliest possible detection, occurring well before the vehicle even enters the turn. The image clearly shows the phantom object appearing as a full obstruction far down the planned trajectory (the light white circle). The system, identifying a high-confidence obstacle directly in its path, immediately terminates the plan and commands a full stop. While an early stop might seem safer, it is still a hazardous phantom braking event. The vehicle stops for no apparent reason long before an intersection, disrupting traffic flow and creating a dangerously unpredictable situation for human drivers who cannot see the phantom obstacle.

**Broader Safety Implications.** This analysis demonstrates




(a) (b) (c)

Figure 12: Autoware's response to phantom detections with the 6-tile mirror at different angles. **(a)** 30°: late-onset phantom triggers immediate emergency stop. **(b)** 45°: earlier detection causes mid-turn braking. **(c)** 60°: early detection triggers pre-turn full stop.

that the vulnerability triggers a spectrum of dangerous *errors of commission*: the vehicle performs unnecessary, hazardous actions ranging from sudden high-$g$ emergency stops to inexplicable halts in live traffic. The severity and timing are directly tunable by the mirror angle parameter.

Although tests were conducted at 8 km/h for safety, autonomous vehicles typically approach intersections at 10–20 km/h, where reduced reaction time and increased momentum amplify the risk of cascade collisions. Crucially, in every case, the Autoware stack performed *exactly as designed*: executing a logical, defensive maneuver based on its sensory input. The vulnerability lies not in the planning software but at the physical-sensor interface, where the vehicle's perception of reality was corrupted by physically legitimate but geometrically deceptive signals. This finding underscores that robust AV safety must extend beyond algorithmic integrity to include defenses against physical-layer sensor deception.

**Note on OR System-Level Evaluation.** For the OR scenario, we did not evaluate full autonomous driving through the removed obstacle, as doing so would require the vehicle to drive into a physical hazard. Instead, we verified the system-level implication at the perception level: Autoware's occupancy grid and planner consistently treated the mirror-occluded region as free space (Figure 11), confirming that the vehicle would proceed into the obstructed zone if operating autonomously. A full collision demonstration for OR would require a simulation-only evaluation with a destructible obstacle, which we leave to future work.

# 6 Physics-Informed Empirical Modeling

The full-stack assessment revealed that even a small number of phantom points can cascade into safety-critical failures. This section derives analytical models that characterize mirror-induced artifacts as functions of controllable physical parameters, formalizing the geometric and statistical mechanisms that govern artifact appearance, depth, and intensity. These models serve two purposes: (1) they provide a quantitative understanding of the vulnerability's underlying physics, and (2) they enable realistic artifact injection in simulation.

## 6.1 Feature Extraction Pipeline

From the experimental data (Section 4), we extracted mirror-induced artifact points through a frame-differencing procedure. For each trial, we aligned the mirror-present scan with a baseline scan captured at the same vehicle position and under identical conditions but without the mirror. Points appearing exclusively in the mirror-present scan were labeled as mirror-induced artifacts. We applied fine-grained spatial registration using the Iterative Closest Point (ICP) algorithm to compensate for minor pose deviations and vehicle motion between runs. The resulting artifact clusters were segmented and annotated with the physical parameters $(d, \theta, A)$: LiDAR-to-mirror distance, mirror tilt angle, and mirror surface area.

From each cluster, we extracted four quantitative features: (1) **Radial distance** ($R_{\text{art}}$): the magnitude of the centroid vector, representing the artifact's apparent distance from the sensor; (2) **Lateral offset** ($X_{\text{art}}$): the displacement of the centroid from the LiDAR's forward axis; (3) **Point count** ($N_{\text{art}}$): the number of points per frame, serving as a proxy for artifact salience; (4) **Appearance probability** ($P_{\text{app}}$): the fraction of frames in which an artifact cluster (with $N > 0$) appears under a given configuration.

## 6.2 Model Formulation

This section models the dependence of artifact characteristics on mirror distance $d$, tilt angle $\theta$, and surface area $A$. We propose four semi-empirical formulations derived from reflection geometry and LiDAR time-of-flight behavior and validate them against the experimental observations in Section 4. All fitted parameters and their physical interpretations are provided in Appendix A.

❶ **Artifact Location Model.** The artifact's perceived position is determined by the two-hop path: the LiDAR pulse strikes the mirror, reflects toward a secondary surface (e.g., wall or ground), bounces back to the mirror, and returns to the sensor. Because the sensor interprets the total round-trip time as a straight-line range, it registers a phantom point at a location with no real object. We derive lateral and radial coordinates separately.

*Lateral Offset.* A planar mirror rotated by a yaw angle $\theta$ relative to the sensor's optical axis deflects the beam laterally

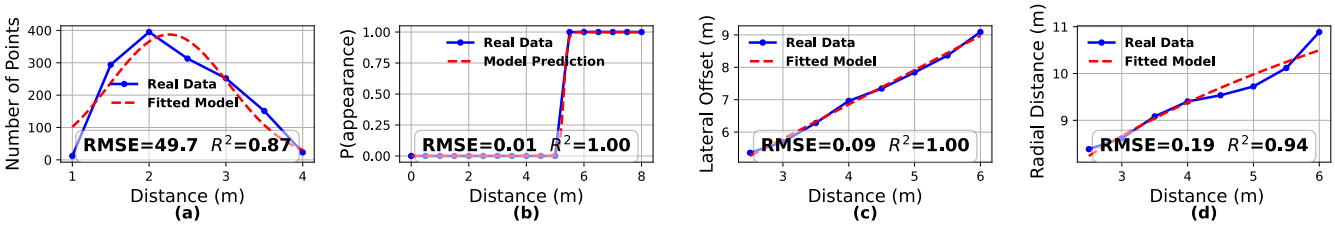



Figure 13: Model validation against experimental data for representative configurations. **(a)** Point count model captures the bell-shaped intensity profile (30°, 0.36 m$^2$). **(b)** Appearance probability model delineates the detection window (60°, 0.6 m$^2$). **(c, d)** Lateral offset and radial distance models track artifact location (45°, 0.36 m$^2$).

by 2θ (law of reflection). This angular deflection produces a lateral displacement that grows with propagation distance. Since the post-reflection path scales approximately with the sensor-to-mirror distance $d$ during vehicle approach, the first-order displacement is:

$$X_{\text{geom}} = d\tan(2\theta). \tag{3}$$

In practice, the sensor and mirror are mounted at different heights, and the beam strikes the mirror as a finite spot rather than an infinitesimal point. The height difference slightly alters the effective deflection geometry (multiplicative correction $c_X$), while the finite beam footprint on the tilted surface introduces a systematic lateral bias (additive shift $\delta_X$):

$$X_{\text{art}}(d,\theta) = c_X\, d\tan(2\theta) + \delta_X. \tag{4}$$

Fitted values ($c_X \approx 0.98$, $\delta_X \approx -0.05$ m) confirm close agreement with the ideal geometry of Equation 3.

*Radial Distance.* The LiDAR computes range as half the total round-trip time multiplied by the speed of light. For the two-hop path, this distance is necessarily longer than the direct round-trip to the mirror. With a perfectly collimated beam, the perceived range would grow linearly with $d$. However, real beams diverge: the beam diameter on the mirror grows as $d\cdot\phi$ (where $\phi$ is the divergence angle), so for a mirror of width $w$ the captured fraction is:

$$f_{\text{capture}}(d) \approx \min\left(1,\ \frac{w}{d\,\phi}\right). \tag{5}$$

As $d$ increases, the beam overfills the mirror and only the central portion is reflected. Central rays travel shorter geometric paths than peripheral ones, so clipping the edges biases the LiDAR's timing estimate (which operates on the peak or leading edge of the return pulse) toward shorter values. This produces sub-linear growth of perceived range:

$$R_{\text{base}}(d) = c_R\, d^{n_d}, \quad n_d < 1. \tag{6}$$

Mirror tilt θ also modifies the total path length by changing how far the deflected beam travels to reach the secondary surface. Expanding this angular dependence to second order gives the full model:

$$R_{\text{art}}(d,\theta) = c_R\, d^{n_d}\,(1 + a_1\theta + a_2\theta^2), \tag{7}$$

where fitting yields $n_d \approx 0.88$, confirming the expected sub-linear regime. Mirror area $A$ has negligible effect on artifact location, consistent with geometric optics: image position depends on angles and distances, not mirror size.

❷ **Artifact Point Count Model.** The LiDAR records a point only when received power exceeds a threshold $P_{\text{min}}$. For a single-hop return at range $R$:

$$P_r = \frac{P_t\,\rho\,A_r}{R^4}\,\eta, \tag{8}$$

where $P_t$ is transmitted power, ρ target reflectivity, $A_r$ receiver aperture, and η system efficiency. In the two-hop case, the effective range, cross-section, and reflectivity become coupled functions of $(d,\theta,A)$, precluding a closed-form solution. We identify three dominant effects and model each independently.

*Mirror area A.* A larger mirror intercepts more of the divergent beam. For a Gaussian beam profile, captured power scales super-linearly when the mirror is smaller than the footprint (the beam center carries disproportionately more energy), motivating a power-law term $A^{\beta}$ with $\beta > 1$.

*Mirror tilt* θ. Tilting reduces the projected area seen by the beam (foreshortening). In our two-hop geometry this loss applies on both incoming/outgoing reflections, and angular misalignment further reduces return coupling, producing a steeper-than-cosine falloff modeled as $\cos^{\gamma}(\theta)$ with $\gamma > 1$.

*Distance d.* The dependence is non-monotonic. At short range, the deflected beam overshoots nearby surfaces and the return angle exceeds the receiver's field of view. At long range, $R^{-4}$ attenuation dominates. Between these regimes, a peak occurs at an optimal distance $\mu(\theta)$. We model this bell-shaped envelope with a Gaussian:

$$g(d,\theta) = \exp\left(-\frac{(d-\mu(\theta))^2}{2\sigma^2}\right), \tag{9}$$

where $\mu(\theta)$ shifts the peak with angle and σ controls the width. Combining the three effects:

$$N_{\text{art}}(d,\theta,A) = c_0 A^{\beta}\cos^{\gamma}(\theta)\exp\left(-\frac{(d-\mu(\theta))^2}{2\sigma^2}\right), \tag{10}$$

with fitted $\beta \approx 1.25$, $\gamma \approx 3.5$.

❸ **Appearance Probability Model.** An artifact is detectable only when (i) the deflected beam actually hits a secondary surface rather than escaping into open space, and (ii) the return signal exceeds $P_{\min}$. At short range, the steep $2\theta$ deflection sends the beam past nearby surfaces, making the two-hop path physically impossible. At long range, the extended multi-leg path causes $1/R^4$ decay to push the signal below the noise floor. These opposing constraints create a finite detection window $[d_{\min}, d_{\max}]$.

The transitions at both boundaries are gradual (the fraction of valid rays and the SNR both degrade smoothly), so we model each edge with a logistic sigmoid and compose them into a bandpass window:

$$P_{\text{app}}(d,\theta,A) = \frac{1}{1+e^{-k(d-d_{\min})}} \cdot \frac{1}{1+e^{k(d-d_{\max})}}, \tag{11}$$

where $k$ controls transition steepness. The boundaries shift with mirror configuration: steeper $\theta$ deflects the beam more aggressively, requiring a greater distance for the geometry to work, pushing $d_{\min}$ outward; larger $A$ captures more energy and extends the detectable range, increasing $d_{\max}$. We model both as linear functions:

$$d_{\min}(\theta,A) = b_0^{\min} + b_1\theta + b_2 A, \tag{12}$$

$$d_{\max}(\theta,A) = b_0^{\max} + c_1\theta + c_2 A. \tag{13}$$

## 6.3 Model Fitting and Validation Summary

Figure 13 presents representative validation results. The lateral offset model (c) achieves a near-perfect fit ($R^2 \approx 0.98$ globally), confirming that the first-order geometric prediction $d \cdot \tan(2\theta)$ accurately describes the sideways displacement. The radial distance model (d) achieves $R^2 \approx 0.92$, with the sub-linear exponent $n_d = 0.88$ consistently capturing the beam-spillover effect across all configurations. The appearance probability model (b) achieves $R^2 \geq 0.87$ across all tested setups, and $R^2 = 1.0$ for several, accurately reproducing the sharp on/off detection windows observed in the data. The point count model (a) achieves a global $R^2 \approx 0.85$; we note that certain configurations (e.g., $30^\circ$, $0.36\,\text{m}^2$) exhibit higher absolute residuals (RMSE $\sim$49.7) driven by the high peak counts at these geometrically favorable setups, but the relative error remains bounded and the model correctly captures the bell-shaped trend in all cases. The complete per-configuration validation across all setups is provided in Appendix A.1, and all fitted parameter values with their physical interpretations are tabulated in Appendix A.

# 7 Simulation-Based Evaluation in CARLA

Here we integrate the previously derived artifact models into CARLA [4] by injecting synthetic LiDAR points whose position and structure follow the predicted artifact behavior. CARLA, like most AV simulators, relies on ray casting that assumes purely diffuse reflections, so mirror-like materials produce no measurement artifacts despite being a known source of errors in real deployments. Consequently, perception stacks tested only in simulation never encounter reflection-induced failures. By generating artifacts according to our empirical formulations, we extend the simulator with realistic specular reflection effects and enable safe, repeatable evaluation of AV robustness under physically consistent conditions.

## 7.1 Injection Pipeline

The pipeline operates in real-time within CARLA's simulation loop and is formalized as a four-stage Algorithm 1:

**Stage 1: State extraction.** The pipeline begins by calculating the precise geometric relationship between the ego vehicle and the adversarial mirror. Using CARLA's APIs, we access the world-space transforms of both the LiDAR sensor and the mirror object. These are used to compute the three critical input parameters for our models: the instantaneous LiDAR-to-mirror distance ($d$), mirror tilt angle ($\theta$), and mirror surface area ($A$). This step effectively translates the dynamic 3D world of the simulator into the specific parametric inputs our empirical models require.

**Stage 2: Probabilistic trigger.** The extracted state vector is passed into a stochastic gate to determine whether an artifact should be injected. The inherent randomness in real-world sensor physics, where signals fluctuate near detection thresholds, is captured by our empirically derived appearance probability $P_{\text{app}}$. For a given state $(d,\theta,A)$, the model outputs the likelihood of artifact presence. A uniformly distributed random number $r \in [0,1]$ is sampled, and the artifact is generated only if $r < P_{\text{app}}(d,\theta,A)$. The probabilistic mechanism reflects the intermittent flicker of marginal detections observed in practice.

**Stage 3: Artifact synthesis.** Once triggered, the pipeline synthesizes a phantom object based entirely on the validated empirical models. All characteristics are derived from the same state vector $(d,\theta,A)$. The point count model $N_{\text{art}}$ determines the number of LiDAR points forming the artifact. Simultaneously, the location models $X_{\text{art}}$ and $R_{\text{art}}$ estimate the 3D centroid $C_{\text{art}}$. A cluster of $N$ points is then sampled from a multivariate Gaussian distribution $\mathcal{N}(C_{\text{art}}, \Sigma_{\text{fixed}})$, where $\Sigma_{\text{fixed}}$ is a small, constant covariance matrix. This process ensures a compact, spatially coherent artifact that resembles a physical object in the scene.

**Stage 4: Point cloud injection.** The final stage performs the insertion of the synthesized artifact cluster, $\mathcal{F}$, into the LiDAR data stream. The generated 3D points are merged with the native LiDAR scan, $\mathcal{L}_{\text{native}}$, using a set union: $\mathcal{L}_{\text{modified}} = \mathcal{L}_{\text{native}} \cup \mathcal{F}$. This modification is applied directly at the sensor-data level, prior to any downstream perception or control processing. As a result, the attack remains entirely transparent to the autonomy stack, which interprets the modified point

**Algorithm 1** Model-Driven Artifact Injection Pipeline

**Require:** Current mirror configuration $(d, \theta, A)$
**Require:** Native LiDAR scan $\mathcal{L}_{\text{native}}$
**Require:** Empirical models: $P_{\text{app}}, N_{\text{art}}, X_{\text{art}}, R_{\text{art}}$

$p_{app} \leftarrow P_{\text{app}}(d, \theta, A)$
**if** r() $< p_{app}$ **then**
  $N \leftarrow \text{int}(N_{\text{art}}(d, \theta, A))$
  $R \leftarrow R_{\text{art}}(d, \theta)$
  $X \leftarrow X_{\text{art}}(d, \theta)$
  $C_{\text{art}} \leftarrow \text{ConvertTo3D}(R, X)$ ▷ Convert to 3D point
  $\mathcal{F} \leftarrow \emptyset$ ▷ Initialize empty artifact cluster
  **for** $i = 1$ to $N$ **do**
    Sample point $\mathbf{p}_i \sim \mathcal{N}(C_{\text{art}}, \Sigma_{\text{fixed}})$
    Add $\mathbf{p}_i$ to $\mathcal{F}$
  **end for**
  $\mathcal{L}_{\text{modified}} \leftarrow \mathcal{L}_{\text{native}} \cup \mathcal{F}$
**else**
  $\mathcal{L}_{\text{modified}} \leftarrow \mathcal{L}_{\text{native}}$
**end if**
**return** Modified LiDAR scan $\mathcal{L}_{\text{modified}}$

cloud as authentic sensor input, enabling the observation of unfiltered system reactions.

## 7.2 Scenario-Based AV Response Analysis

The system-level impact of our attack was evaluated in a critical urban driving scenario designed in CARLA. This scenario enables observation of the complete event chain, from initial perception failure to a physical collision, thereby providing a clear measure of the threat's severity.

**Simulation Setup.** Our primary scenario involves two vehicles approaching a T-intersection. The lead vehicle, which we term the "Ego AV," is equipped with our full attack injection pipeline and a standard autonomous driving agent (`BehaviorAgent` in CARLA). A second "Follower Vehicle" trails the Ego AV, governed by a basic car-following logic that maintains a safe distance based on its own perception. An adversarial mirror is placed at the corner of the T-intersection, positioned to trigger model-based artifact injection that simulates a phantom object in the Ego AV's path.

**Qualitative Outcome.** As the Ego AV approaches the mirror, the pipeline continuously evaluates the configuration state $(d, \theta, A)$ and determines whether artifact generation should occur. Once the appearance probability exceeds the stochastic threshold, a dense cluster of phantom points is injected into the Ego AV's LiDAR stream at a physically consistent location and depth. The perception module interprets this cluster as a nearby obstacle occupying the driving corridor, which propagates to tracking and planning as a potential collision risk. The safety layer therefore overrides the nominal trajectory and initiates emergency braking. The follower vehicle reacts only after observing the abrupt deceleration of the Ego AV and cannot brake in time, resulting in a rear-end collision (Figure 14).

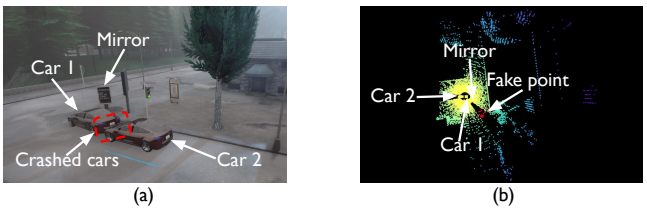



Figure 14: Cascade collision in CARLA. The Ego AV (Car 1) emergency-brakes for a phantom obstacle; the Follower (Car 2) rear-ends it. **(a)** Collision moment. **(b)** Ego AV's LiDAR view showing injected phantom points.

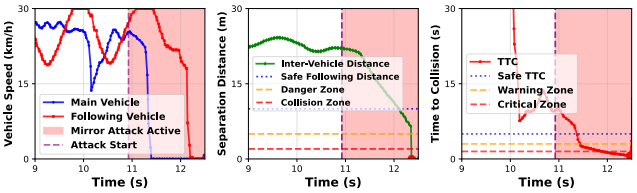



Figure 15: System-level dynamics following artifact injection at $t \approx 11.0$ s. Ego AV emergency brakes; Follower Vehicle fails to react, leading to collision (separation and TTC reach zero).

**Quantitative Analysis.** Figure 15 presents time-aligned vehicle dynamics during a representative scenario. At $t \approx 11.0$ s, artifact injection begins. The Ego AV (blue) decelerates from $>25$ km/h to zero within $\sim$1 s. The Follower (red) continues at speed before reacting with a delay, causing the separation distance to collapse to zero. The Time-to-Collision (TTC) drops below the critical 3-second threshold and ultimately reaches zero, confirming a collision. This causal chain, perception deception $\rightarrow$ emergency braking $\rightarrow$ rear-end collision, demonstrates how a passive optical manipulation at the sensor layer can cascade into a catastrophic system-level failure.

**Point Count and Attack Success.** Table 2 summarizes the relationship between mirror configuration, phantom point count, perception output, and system-level behavior across both the real-world Autoware experiments (Section 5) and the closed-loop CARLA evaluation presented above. The three mirror areas span roughly an order of magnitude in artifact density (from $\approx$25 to $\approx$200 points per frame), yet all configurations produce the same system-level outcome: emergency braking followed by rear-end collision in simulation, and planner-level obstacle response in the real-world perception traces. This reveals a two-tier threshold structure. Dense clusters ($\gtrsim$50 points) drive high-confidence semantic misclassification as "Vehicle" (65% and 74% confidence), whereas sparse clusters ($\approx$25 points) are labeled "UNKNOWN" but still populate occupancy cells and trigger defensive planner behavior. The operationally relevant threshold for attack success is therefore the occupancy-grid threshold, not the semantic-classification threshold, and it is met even by the sparsest configurations we

| Area (tiles) | Pts/Frame | Perception | Outcome |
|---|---|---|---|
| 0.18 $m^2$ (2) | ≈25 | UNKNOWN | Brake + Crash |
| 0.36 $m^2$ (4) | ≈55 | CAR (65%) | Brake + Crash |
| 0.60 $m^2$ (6) | ≈200 | CAR (74%) | Brake + Crash |

Table 2: Point count and attack outcomes across mirror configurations, real-world perception (Section 5), and CARLA closed-loop behavior.

tested. This indicates that the attack does not require finely tuned geometry or large mirrors to produce unsafe system-level behavior.

**Speed sensitivity.** Since varying vehicle speed in our real-world test environment was not feasible for safety reasons, we used CARLA to evaluate the vulnerability at realistic urban approach speeds of 25, 30, and 35 km/h. The injected phantom triggered emergency braking and a rear-end collision at all three speeds, with collision severity increasing at higher velocities due to greater closing momentum and shorter available braking margins.

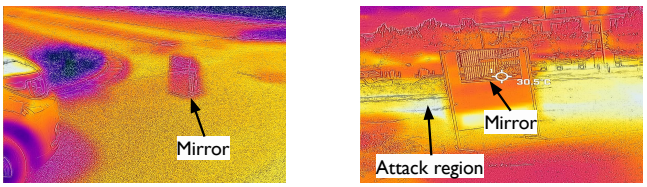
## 8 Discussion and Countermeasures

Our findings show that mirror-induced artifacts exploit fundamental assumptions in LiDAR perception, creating serious safety risks for autonomous systems. This section examines potential defense strategies and discusses their limitations.

### 8.1 Defense Strategies

① **Thermal Camera Augmentation.** The fundamental principle of this defense is to leverage a different sensing modality that is not susceptible to the same physical deception. Thermal (long-wave infrared) cameras detect emitted heat rather than reflected light. In OA, the phantom object is a LiDAR-specific illusion that lacks physical substance and therefore has no corresponding heat source. A thermal camera observing the phantom's perceived location would see only the ambient background temperature, directly falsifying the LiDAR detection. In OR, the mirror itself becomes the detectable anomaly. Its surface typically presents a thermal signature distinct from the surrounding road, creating a clear thermal discontinuity that reveals the presence of the adversarial element. In both cases, the thermal modality provides critical information to either invalidate a phantom or detect the tool of deception.

**Implementation and Evaluation.** We integrated a FLIR One thermal camera into our test vehicle and developed a detection module to identify anomalous thermal regions. The camera proved effective in both scenarios (see Figure 17). In the OR case (**a**), the mirror panel is clearly delineated as a cool (purple) rectangular surface against the warmer (yellow) asphalt, creating a distinct thermal boundary that allows for direct detection of the adversarial element. In the OA scenario (**b**), the thermal image shows that the region where LiDAR detected



(a) (b)

Figure 16: Thermal imagery of mirror attack scenarios. **(a)** OR: The mirror is visible as a cool (purple) surface against the warmer (yellow) road. **(b)** OA: The phantom object region shows no thermal anomaly, matching the ground temperature and confirming it is not a real object.

a phantom vehicle is thermally uniform with its surroundings. The absence of any localized heat source different from the ground plane confirms the object's non-physical nature.

**Limitations.** Despite its promise, thermal imaging is not a panacea. Its efficacy can be reduced by environmental factors, such as high ambient temperatures that minimize the thermal contrast between objects (thermal crossover). The resolution of thermal cameras is typically lower than that of RGB cameras, making it difficult to detect small or distant mirrors. Furthermore, practical deployment involves challenges such as cost, calibration, and maintenance, as lenses can be obscured by rain or dirt.

② **Multi-Sensor Fusion.** Another potential defense strategy involves leveraging the inherent redundancy of an autonomous vehicle's multi-modal sensor suite. Standard sensor fusion algorithms typically operate on the principle of data reinforcement, assuming that different sensors will provide consistent, corroborating evidence about the state of the world. Mirror-based attacks are designed to violate this assumption by creating a fundamental *disagreement* between modalities. An OA creates a phantom visible only to LiDAR, while an OR can make a real object invisible only to LiDAR. A defense can therefore be built around a "fusion consistency monitor" designed to detect and flag these specific inter-sensor conflicts as anomalies.

**Conceptual Implementation and Mitigation.** This defense would be implemented as a logic layer that continuously cross-verifies detections between LiDAR and other sensors, such as RGB cameras or radar. Aligned with proposals such as that of [15], the system would monitor for high-confidence inconsistencies. For instance, in an OA scenario, a LiDAR-detected object that lacks a corresponding visual object in the camera feed would be flagged as anomalous. Conversely, in an OR scenario, a visually-detected obstacle that corresponds to a region of navigable free space in the LiDAR data would trigger a similar alert. Upon detecting such a conflict, the system could respond by significantly lowering the confidence score of the suspect perception, or by triggering a safe-state behavior, such as reducing speed, until the conflict is resolved by subsequent sensor readings.

**Limitations and Challenges.** The primary challenge of this approach lies in robustly disambiguating malicious inconsistencies from benign, modality-specific sensor failures. A naively implemented consistency check could generate high-risk false alarms. For example, in low-light conditions, a camera will naturally fail to detect a real obstacle that is clearly visible to LiDAR; a simple check would incorrectly discard this valid detection as a phantom. Similarly, benign physical phenomena, such as a retro-reflective road sign being highly salient to a camera but producing sparse returns for LiDAR, can also create disagreement. Therefore, developing a sophisticated consistency model that accounts for varying environmental conditions and the unique operational characteristics of each sensor without introducing an unacceptable rate of false positives is a non-trivial research challenge. Other advanced concepts, such as using redundant LiDARs at different physical locations to break the attack's geometric constraints, also warrant future exploration.

③ **Intensity-Based Anomaly Detection.** A recent line of work leverages LiDAR return intensity to detect physical adversarial attacks, observing that injected points often exhibit abnormal intensity relative to genuine returns [17]. This defense is effective against active laser spoofing, where received power is set by the attacker's emitter and does not match the expected intensity–range relationship. Mirror artifacts, however, are the vehicle's own pulses following a legitimate two-hop path: the received power follows Equation 8 applied to the total path length $d_{LM} + d_{MS}$, which is also the range the sensor reports. Intensity and range are therefore physically consistent by construction. The sensor observes exactly the return it would expect from a real object at that range with the reflectivity of the secondary surface, leaving no radiometric anomaly to detect.

**Empirical Evaluation.** To test this quantitatively, we computed the distribution of calibrated Ouster reflectivity values (0 to 255 scale) for mirror-induced artifact points and legitimate returns from the same scans, across three representative configurations spanning our parameter grid: (45°, 0.36 m$^2$), (30°, 0.60 m$^2$), and (60°, 0.18 m$^2$). For each configuration, we compared ∼1 to 3 k artifact points against ∼3 M legitimate returns drawn scene-wide, excluding only the artifact region, since a per-point detector operates across the full scan. Table 3 reports the key detection metrics.

The detector operating in the Park direction, flagging high-reflectivity points as adversarial, performs worse than random, with AUC values of 0.12 to 0.17. This reflects a physical reality: the two-hop path attenuates the return through mirror reflectivity (∼0.95), the diffuse albedo of the secondary surface (∼0.01 to 0.10), and additional geometric spreading loss. Mirror artifacts are therefore systematically *dimmer* than typical direct returns, the opposite of active-spoofing injections. At a deployable 5% false-positive rate, the Park-style detector catches only 0.7 to 4.4% of artifact points. We also evaluated the inverted detector, flagging low-reflectivity points. While this captures the directional asymmetry of the data, it still fails at operationally relevant thresholds: catching 95% of artifacts requires flagging 52 to 86% of legitimate returns, because dark asphalt, shadowed regions, and low-albedo road surfaces populate the same low-reflectivity tail as mirror artifacts. Neither threshold direction yields a deployable operating point, confirming that intensity-based detection cannot address this class of vulnerability.

| Config | $\mu_{\text{art}}$ | $\mu_{\text{base}}$ | AUC | TPR @ 5% FPR |
|---|---|---|---|---|
| (45°, 0.36 m$^2$) | 3.1 | 12.9 | 0.12 | 0.7% |
| (30°, 0.60 m$^2$) | 4.1 | 12.8 | 0.13 | 1.6% |
| (60°, 0.18 m$^2$) | 10.4 | 12.7 | 0.17 | 4.4% |

Table 3: Intensity-based detection performance against mirror artifacts. AUC values well below 0.5 indicate that the Park-direction detector (flag high-reflectivity outliers) is anti-correlated with the attack; artifacts are dimmer than legitimate returns. At a deployable 5% false-positive budget, the detector catches essentially none of the attack.

## 8.2 Broader Implications

### 8.2.1 Generality Across LiDAR Platforms

We replicated the Object Addition experiment on four additional LiDAR sensors to verify that the attack is not specific to our test platform. The sensors span three vendors and a 4× range in vertical resolution: Ouster OS1-64, Ouster OS0-64 (wide vertical FOV variant), Robosense RS-Helios 5515 (32 channels), and Hesai XT32M2X (32 channels). Each sensor was deployed in a static outdoor configuration at a representative mirror tilt of 45°, matching the geometry used in our main experiments. We did not integrate these sensors into the autonomous vehicle stack, as physical integration and per-sensor calibration are non-trivial. The goal is to confirm reproducibility of the underlying physical phenomenon rather than to repeat the full perception-and-planning analysis. Figure 17 compares the baseline scans without the mirror against the corresponding mirror-present scans for each sensor. Across all four platforms, the mirror induces a spatially coherent phantom cluster behind the mirror surface, consistent with the two-hop reflection mechanism described in Section 2. These results indicate that the phenomenon is not an artifact of a single sensor model, but rather a cross-platform vulnerability that can manifest across LiDAR configurations representative of those used in autonomous driving systems.

### 8.2.2 Geometric Trust as a Security Assumption

Our results show that the vulnerability does not stem from noise, learning errors, or injected signals, but from a deeper architectural assumption: LiDAR perception treats measured range as geometrically correct by construction. The autonomy stack verifies signal reception but never verifies whether

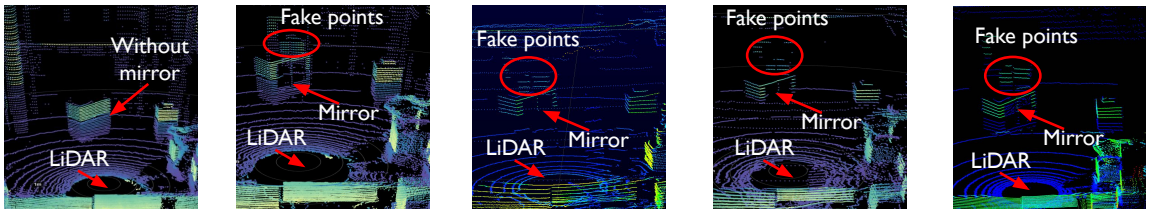



Figure 17: Object Addition across four LiDAR platforms. **(a)** Baseline (no mirror). **(b)** Ouster OS1-64. **(c)** Robosense RS-Helios 5515. **(d)** Ouster OS0-64. **(e)** Hesai XT32M2X.

the light propagation path is physically plausible. Specular reflections preserve valid timing and intensity statistics while altering spatial origin, producing observations that are internally consistent yet geometrically incorrect.

This distinguishes mirror-induced artifacts from traditional spoofing in a way that has direct defensive implications. Active attacks introduce foreign signals and can in principle be detected through authentication or waveform analysis. Mirror artifacts, by contrast, are the sensor's own photons returning through a physically valid but geometrically unexpected path. The deception is entirely in the propagation domain, not the signal domain. Consequently, no signal-level defense can address this class of vulnerability. Effective countermeasures must instead reason about *path plausibility*: whether the spatial configuration of observed returns is consistent with single-bounce propagation. This represents a verification layer that no current production LiDAR system or perception stack provides.

### 8.2.3 Implications for AV Validation

Current AV validation practices rely heavily on simulation and curated datasets that implicitly assume direct line-of-sight propagation. Perception systems may therefore satisfy standard safety metrics while never encountering geometrically inconsistent measurements. The absence of such scenarios represents a systematic blind spot in testing methodology rather than an implementation flaw.

By introducing physically calibrated multi-path artifacts into simulation, our framework exposes failure modes that remain invisible under conventional evaluation. More broadly, this suggests that validating autonomous systems requires explicitly modeling sensing physics, including multi-path propagation and specular surface interactions, not only sensor accuracy or algorithmic performance.

## 9 Limitations and Future Work

**Sensor Generality.** Our experiments used a spinning ToF LiDAR (Ouster OS1-128). While the underlying specular reflection vulnerability applies to any ToF system, Frequency Modulated Continuous Wave (FMCW) LiDAR and sensors with full-waveform analysis capabilities may exhibit different susceptibility profiles. Validation across sensor technologies is an important next step.

**Environmental Conditions.** All tests were conducted in clear, dry weather. Rain, fog, and humidity can alter mirror reflectivity and beam propagation, potentially affecting artifact characteristics in ways our current models do not capture. Systematic investigation of weather effects is needed.

**Vehicle Speed.** Our real-world experiments were limited to $\leq$8 km/h for safety. At higher urban speeds (30–50 km/h), the reduced number of frames within the appearance window could affect tracking-module behavior. However, our models predict that the appearance window spans several meters of distance even at 60°, and safety-critical planners are designed to react to even brief high-confidence detections. Controlled high-speed testing remains a priority.

**Defense Integration.** The countermeasures discussed in Section 8 were evaluated at a conceptual or prototype level. Full integration into an operational AV stack with rigorous false-positive analysis is needed before they can be considered practical defenses.

## 10 Related Work

**Active LiDAR Spoofing.** Most known LiDAR spoofing attacks use external laser emitters to inject fake pulses, creating phantom objects at arbitrary distances [2, 5, 7, 10, 18, 19]. These attacks are effective but require specialized hardware, line-of-sight access, and often precise synchronization with the victim sensor's timing. Our work demonstrates that comparable perception failures can be induced passively, without any electronic emissions.

**Passive Physical-World Attacks.** Passive approaches have used 3D-printed adversarial objects [1, 24] or crafted road patterns [27] to exploit learned perception models. These target model-specific vulnerabilities rather than the fundamental sensing modality. The "Shadow Hack" work [11, 12] demonstrated point removal using reflective mats but did not explore object addition or derive predictive models. Other work examined adversarial materials on road signs [31] or used drones for occlusion [32]. Our contribution differs in providing a comprehensive analysis of both addition and removal, grounded in geometric optics modeling, with end-to-end AV

stack evaluation.

**LiDAR Artifacts and Reflective Surfaces.** The robotics community has long recognized challenges from reflective surfaces in LiDAR mapping [3, 23, 28], and recent work has proposed SLAM algorithms robust to such artifacts [9, 13, 25]. However, these studies treat reflections as incidental noise rather than adversarial threats, and none provide the empirical models necessary for systematic evaluation.

**Defenses and Sensor Fusion.** Defenses against LiDAR spoofing include statistical filtering [8, 22], signal-level authentication, and multi-sensor fusion [1, 6, 15]. These approaches assume that spoofed signals differ from legitimate returns in detectable ways. Mirror-based artifacts challenge this assumption: they are generated from the vehicle's own pulses and produce physically legitimate returns, making them fundamentally harder to distinguish from benign reflections.

## 11 Conclusion

This work presents the first systematic vulnerability analysis of LiDAR-based autonomous driving systems against mirror-induced perception failures. Through controlled real-world experiments, we demonstrate that inexpensive planar mirrors can corrupt LiDAR point clouds to create phantom obstacles or erase real ones. We trace these artifacts through the Autoware autonomy pipeline, showing they trigger high-confidence false detections (up to 74% "Vehicle" confidence), corrupt occupancy maps, and cause safety-critical failures including emergency braking and cascade collisions.

Our physics-informed empirical models capture the geometric and radiometric properties of mirror-induced artifacts with high fidelity, enabling their integration into a CARLA-based simulation framework that addresses a gap in current simulators' inability to model specular reflection phenomena. Both the real-world validation and simulation framework provide a comprehensive methodology and dataset to support the development of reflection-aware perception systems.

The vulnerability we characterize is particularly challenging to address because mirror-based artifacts are generated from the vehicle's own signals, making them physically indistinguishable from legitimate returns at the signal level. Our analysis of potential countermeasures, including thermal imaging and multi-sensor fusion, reveals that each faces significant practical limitations. Developing robust defenses against physically legitimate but geometrically deceptive sensor inputs remains a critical open challenge for autonomous vehicle safety.

## Acknowledgments

This work was supported in part by the DePhaSe-CPS and CORTESE projects, and by the CYPHER-AV project funded by the German Federal Ministry of Transport. The authors also acknowledge the FASTer project (EFRE-HS-0000774) and the htw saar ITS Research Group, and thank Aymen Bouferroum and Jishitha Kondaveti for their support.

## References

[1] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *2021 IEEE symposium on security and privacy (SP)*, pages 176–194. IEEE, 2021.

[2] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 2267–2281, 2019.

[3] Deeptha Damodaran, Saeed Mozaffari, Shahpour Alirezaee, and Mohammed Jalal Ahamed. Experimental analysis of the behavior of mirror-like objects in lidar-based robot navigation. *Applied Sciences*, 13(5):2908, 2023.

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[5] Masashi Fukunaga and Takeshi Sugawara. Random spoofing attack against lidarbased scan matching slam. *VehicleSec2024*, 2024.

[6] R Spencer Hallyburton, Yupei Liu, Yulong Cao, Z Morley Mao, and Miroslav Pajic. Security analysis of {Camera-LiDAR} fusion against {Black-Box} attacks on autonomous vehicles. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1903–1920, 2022.

[7] Zhongyuan Hau, Kenneth T Co, Soteris Demetriou, and Emil C Lupu. Object removal attacks on lidar-based 3d object detectors. *arXiv preprint arXiv:2102.03722*, 2021.

[8] Zhongyuan Hau, Soteris Demetriou, Luis Muñoz-González, and Emil C Lupu. Shadow-catcher: Looking into shadows to detect ghost objects in autonomous vehicle 3d sensing. In *Computer Security–ESORICS 2021: 26th European Symposium on Research in Computer Security, Darmstadt, Germany, October 4–8, 2021, Proceedings, Part I 26*, pages 691–711. Springer, 2021.

[9] Connor Henley, Siddharth Somasundaram, Joseph Hollmann, and Ramesh Raskar. Detection and mapping of specular surfaces using multibounce lidar returns. *Optics Express*, 31(4):6370–6388, 2023.

[10] Zizhi Jin, Xiaoyu Ji, Yushi Cheng, Bo Yang, Chen Yan, and Wenyuan Xu. Pla-lidar: Physical laser attacks against lidar-based 3d object detection in autonomous vehicle. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1822–1839. IEEE, 2023.

[11] Ryunosuke Kobayashi, Kazuki Nomoto, Yuna Tanaka, Go Tsuruoka, and Tatsuya Mori. Wip: Shadow hack: Adversarial shadow attack against lidar object detection. In *Proc. Symp. Veh. Security Privacy (VehicleSec)*, pages 1–7, 2024.

[12] Ryunosuke Kobayashi, Kazuki Nomoto, Yuna Tanaka, Go Tsuruoka, and Tatsuya Mori. Invisible but detected: physical adversarial shadow attack and defense on lidar object detection. In *Proceedings of the 34th USENIX Conference on Security Symposium*, SEC '25, USA, 2025. USENIX Association.

[13] Rainer Koch, Stefan May, and Andreas Nüchter. Detection and purging of specular reflective and transparent object influences in 3d range measurements. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:377–384, 2017.

[14] Pui Yee Leong and Nur Syazreen Ahmad. Lidar-based obstacle avoidance with autonomous vehicles: A comprehensive review. *IEEE Access*, 2024.

[15] Jinshan Liu and Jung-Min Park. "seeing is not always believing": detecting perception error attacks against autonomous vehicles. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2209–2223, 2021.

[16] Ouster, Inc. OS1 Mid-Range Lidar Sensor. https://ouster.com/products/hardware/os1-lidar-sensor. Accessed: 2026-02-23.

[17] Yeji Park, Hyunsu Cho, Dong Hoon Lee, and Wonsuk Choi. Leveraging intensity as a new feature to detect physical adversarial attacks against lidars. In *2024 Annual Computer Security Applications Conference (ACSAC)*, pages 1002–1014. IEEE, 2024.

[18] Jonathan Petit, Bas Stottelaar, Michael Feiri, and Frank Kargl. Remote attacks on automated vehicles sensors: Experiments on camera and lidar. *Black Hat Europe*, 11(2015):995, 2015.

[19] Takami Sato, Yuki Hayakawa, Ryo Suzuki, Yohsuke Shiiki, Kentaro Yoshioka, and Qi Alfred Chen. LiDAR spoofing meets the new-gen: Capability improvements, broken assumptions, and new attack strategies. *arXiv preprint arXiv:2303.10555*, 2023.

[20] Takami Sato, Ryo Suzuki, Yuki Hayakawa, Kazuma Ikeda, Ozora Sako, Rokuto Nagata, Ryo Yoshida, Qi Alfred Chen, and Kentaro Yoshioka. On the realism of lidar spoofing attacks against autonomous driving vehicle at high speed and long distance. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2025.

[21] Hocheol Shin, Dohyun Kim, Yujin Kwon, and Yongdae Kim. Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications. In *Cryptographic Hardware and Embedded Systems–CHES 2017: 19th International Conference, Taipei, Taiwan, September 25-28, 2017, Proceedings*, pages 445–467. Springer, 2017.

[22] Jiachen Sun, Yulong Cao, Qi Alfred Chen, and Z Morley Mao. Towards robust {LiDAR-based} perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 877–894, 2020.

[23] Haileleol Tibebu, Jamie Roche, Varuna De Silva, and Ahmet Kondoz. Lidar-based glass detection for improved occupancy grid mapping. *Sensors*, 21(7):2263, 2021.

[24] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13716–13725, 2020.

[25] Miguel A Vega-Torres, Alexander Braun, and André Borrmann. Slam2ref: advancing long-term mapping with 3d lidar and reference map integration for precise 6-dof trajectory estimation and map extension. *Construction Robotics*, 8(2):13, 2024.

[26] Jian Wang, Fan Li, Xuchong Zhang, and Hongbin Sun. Adversarial obstacle generation against lidar-based 3d object detection. *IEEE Transactions on Multimedia*, 2023.

[27] Kaichen Yang, Tzungyu Tsai, Honggang Yu, Max Panoff, Tsung-Yi Ho, and Yier Jin. Robust roadside physical adversarial attack against deep learning in lidar perception modules. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 349–362, 2021.

[28] Shih-Wei Yang and Chih-Chung Wang. On Solving Mirror Reflection in LIDAR Sensing. *IEEE/ASME Transactions on Mechatronics*, 16(2):255–265, 2011.

[29] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Centerpoint: Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11784–11793, 2021.

[30] Yongjun Zhang, Pengcheng Shi, and Jiayuan Li. Lidar-based place recognition for autonomous driving: A survey. *ACM Computing Surveys*, 57(4):1–36, 2024.

[31] Shenchen Zhu, Yue Zhao, Kai Chen, Bo Wang, Hualong Ma, and Cheng'an Wei. {AE-Morpher}: Improve physical robustness of adversarial objects against {LiDAR-based} detectors via object reconstruction. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 7339–7356, 2024.

[32] Yi Zhu, Chenglin Miao, Tianhang Zheng, Foad Hajiaghajani, Lu Su, and Chunming Qiao. Can we use arbitrary objects to attack lidar perception in autonomous driving? In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 1945–1960, 2021.

# A Empirical Model Validation and Parameter Analysis

This appendix summarizes the empirical model parameters used in this study. Table 4 reports both the fitted values and their compact physical interpretation for the four empirical models. The coefficients were obtained through global nonlinear least-squares fitting across all relevant experimental configurations and were used to generate the "Fitted Model" curves in our results.

## A.1 Validation Across Experimental Configurations

❶ **Artifact Point Count ($N_{\text{art}}$).** The Gaussian-like model captures the rise-and-fall behavior of phantom point counts across configurations, with $R^2$ values between 0.70 and 1.00 (Figure 18). Point count increases as the mirror subtends a larger solid angle, peaks at favorable alignment, and then decreases as the reflection geometry degrades. Mirror area is the main driver of artifact magnitude, while larger tilt angles reduce point density through foreshortening and less efficient reflection paths.

❷ **Appearance Probability ($P_{\text{app}}$).** The double-sigmoid model fits the observed appearance windows with high accuracy, with $R^2$ values above 0.87 and often close to 1.00 (Figure 19). The results confirm that artifacts appear only within a finite distance interval bounded by $d_{\min}$ and $d_{\max}$.

Table 4: Empirical model parameters.

| Model | Param. | Value | Meaning |
|---|---|---|---|
| Point count $N_{\text{art}}$ $R^2\approx0.85$ | $c_0$ | 2500.0 | Max. point-count scale |
| | $\beta$ | 1.25 | Mirror-area sensitivity |
| | $\gamma$ | 3.5 | Tilt sensitivity |
| | $\mu$ | 3.0 m | Peak-response distance |
| | $\sigma$ | 1.5 m | Effective range width |
| Appearance $P_{\text{app}}$ | $k$ | 15.0 | Window sharpness |
| | $b_0^{\min}$ | $-2.858$ | Lower-bound offset |
| | $b_1, c_1$ | 0.1389 | Angle-boundary shift |
| | $b_2$ | $-0.3003$ | Lower-bound area effect |
| | $b_0^{\max}$ | $-0.946$ | Upper-bound offset |
| | $c_2$ | 1.5390 | Upper-bound area effect |
| Lateral $X_{\text{art}}$ $R^2\approx0.98$ | $c_X$ | 0.98 | Geometry scale |
| | $\delta_X$ | $-0.05$ m | Setup bias |
| Radial $R_{\text{art}}$ $R^2\approx0.98$ | $c_R$ | 1.02 | Distance scale |
| | $n_d$ | 0.88 | Growth exponent |
| | $a_1$ | 0.15 | Linear angle correction |
| | $a_2$ | 0.08 | Quadratic angle correction |

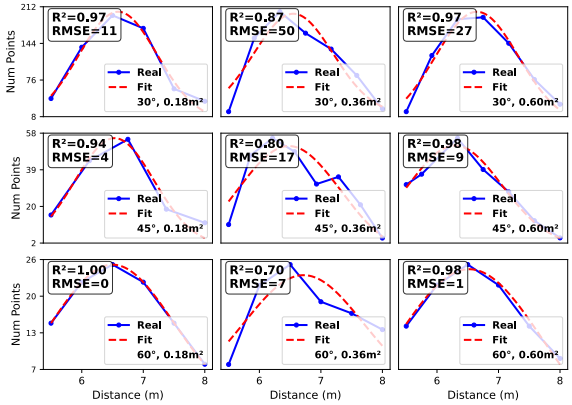



Figure 18: Artifact point count: experimental data ("Real", blue) vs. model predictions ("Fit", red dashed) across all 9 configurations ($\theta \times A$).

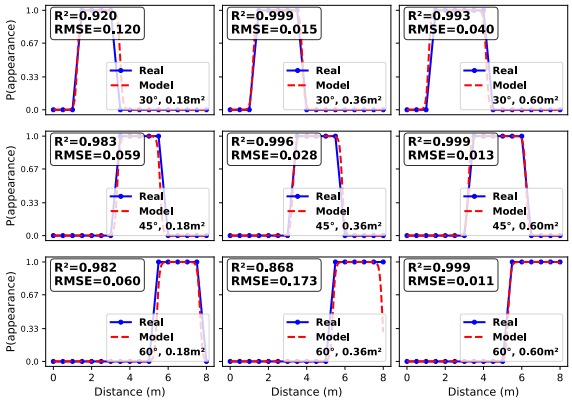



Figure 19: Appearance probability: experimental data ("Real") vs. model predictions ("Model") across all 9 configurations.

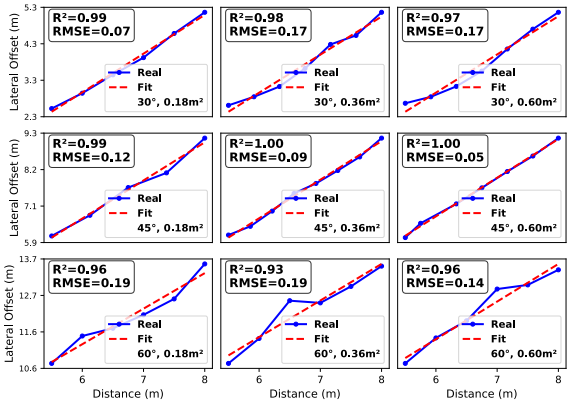



Figure 20: Lateral offset: experimental data ("Real") vs. model predictions ("Fit") across all 9 configurations.

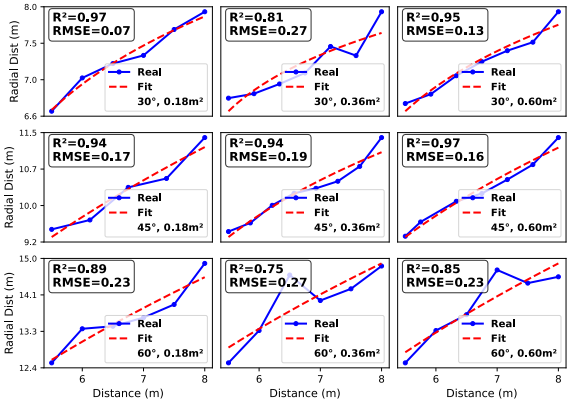



Figure 21: Radial distance: experimental data ("Real") vs. model predictions ("Fit") across all 9 configurations.

Larger mirror areas widen this interval, whereas tilt primarily shifts it along the distance axis.

❸ **Lateral Offset ($X_{\text{art}}$).** The linear model achieves near-perfect agreement, with $R^2 \geq 0.93$ across configurations (Figure 20). The offset scales linearly with distance through the $\tan(2\theta)$ term. Thus, mirror tilt governs phantom placement, while mirror area mainly affects cluster density rather than location.

❹ **Radial Distance ($R_{\text{art}}$).** The power-law model provides strong fits, with $R^2$ values mostly above 0.85 (Figure 21). The perceived radial distance grows nearly linearly with the true distance, while the $d^{n_d}$ term captures sub-linear effects due to beam spillover. Tilt changes the reflected path length, whereas mirror area has little influence on perceived distance.

# B Supplementary Point Cloud Visualizations

This appendix presents supplementary visualizations of the raw LiDAR point clouds generated under different OA and OR configurations. These provide direct visual evidence of how physical parameters influence raw sensor data.

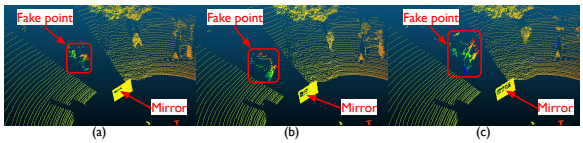



Figure 22: Impact of mirror surface area on phantom point cloud: **(a)** Two tiles (0.18 $m^2$); **(b)** Four tiles (0.36 $m^2$); **(c)** Six tiles (0.60 $m^2$).

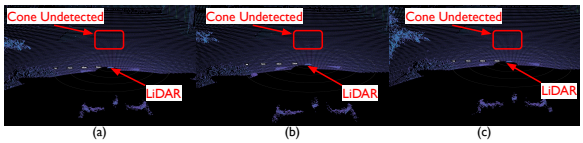



Figure 23: LiDAR point clouds under OR at various mirror angles. **(a)** 15°: cone fully invisible. **(b)** 30°: cone fully invisible. **(c)** 45°: cone fully invisible. The consistency across angles demonstrates a wide operational envelope.

## B.1 OA: Phantom Point Cloud Characteristics

① **Impact of Mirror Surface Area.** Figure 22 shows that larger mirror areas produce denser and more coherent phantom clusters. With 2 tiles (0.18 $m^2$), the cluster survives filtering and is labeled "UNKNOWN." With 4 tiles (0.36 $m^2$), it gains sufficient volumetric structure to be classified as "CAR" with 65% confidence, increasing to 74% with 6 tiles (0.60 $m^2$). Thus, reflective area directly controls deception severity, from a low-level anomaly to confident semantic misclassification.

② **Impact of Mirror Tilt Angle.** Mirror tilt controls both the perceived location and temporal stability of the phantom. At 30°, the restrictive geometry yields a late-onset phantom visible only at close range, while 60° creates a persistent barrier that appears earlier and remains stable over a longer trajectory. The 45° case shows stronger frame-to-frame flicker, indicating that tilt can tune not only phantom position but also temporal consistency.

## B.2 OR: Robustness Across Angles

Figure 23 shows that OR remains effective across mirror angles of 15°, 30°, and 45°. In all cases, the traffic cone disappears from the LiDAR scan while the deflected beams generate valid-looking ground returns, making the obstacle region appear navigable. This consistency indicates that OR does not require precise alignment and can operate over a broad range of practical configurations.