

Defending Autonomous Driving Perception against Adversarial Object-Based Attacks via Motion Planning

Zihao Liu
Iowa State University
Ames, Iowa, USA
zihaliu@iastate.edu

Yan Zhang
Iowa State University
Ames, Iowa, USA
yanzh@iastate.edu

Yi Zhu
Wayne State University
Detroit, Michigan, USA
yzhu39@wayne.edu

Lu Su
Purdue University
West Lafayette, Indiana, USA
lusu@purdue.edu

Chunming Qiao
SUNY at Buffalo
Buffalo, New York, USA
qiao@buffalo.edu

Chenglin Miao
Iowa State University
Ames, Iowa, USA
cmiao@iastate.edu

Abstract

Autonomous vehicles (AVs) rely on perception systems to detect surrounding objects using sensors such as cameras, LiDAR (Light Detection and Ranging), and millimeter-wave (mmWave) radar. However, recent studies have shown that attackers can deceive these systems by strategically placing adversarial objects (e.g., color patches, cardboard, or metal foil) in the driving environment. These attacks pose serious safety risks, yet existing defenses primarily focus on individual sensor modalities and lack generalizability across different sensing systems. To address this gap, we propose the first generalized defense mechanism capable of mitigating various attacks using adversarial objects. Our approach integrates real-time attack detection with trajectory adaptation, guiding the victim AV to positions where the attack is less effective. The defense mechanism combines a deep reinforcement learning (DRL)-based motion planning model, which dynamically adjusts the AV's trajectory, with an uncertainty-aware filtering scheme that refines perception outputs to enhance detection robustness. Extensive experiments in both simulated and real-world environments demonstrate that our defense mechanism effectively mitigates adversarial object-based attacks across different sensing modalities and sensor fusion while maintaining safe and smooth driving behavior.

CCS Concepts

• **Security and privacy** → **Domain-specific security and privacy architectures**; • **Computer systems organization** → **Embedded and cyber-physical systems**.

Keywords

Autonomous driving perception, Adversarial attacks, Defense, Motion planning

ACM Reference Format:

Zihao Liu, Yan Zhang, Yi Zhu, Lu Su, Chunming Qiao, and Chenglin Miao. 2026. Defending Autonomous Driving Perception against Adversarial Object-Based Attacks via Motion Planning. In *ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems (SenSys '26)*, May 11–14, 2026, Saint Malo, France. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3774906.3800480>

1 Introduction

Autonomous driving has made significant advancements in recent years, bringing the vision of fully autonomous vehicles (AVs) closer to reality. AVs rely on advanced perception systems to detect objects in the surrounding environment using onboard sensors such as cameras, LiDAR, and mmWave radar. As a critical component of autonomous driving, the perception system plays a pivotal role in ensuring driving safety by providing accurate environmental awareness to support downstream decision making and motion planning. Recent breakthroughs in deep neural networks (DNNs) have enabled perception modules to achieve high detection accuracy, further strengthening AVs' capability to navigate complex real-world scenarios.

However, the increasing integration of AVs into daily life has incentivized malicious parties to exploit their perception systems, aiming to disrupt normal operations and compromise road safety. Recently, a growing body of research has focused on attacking AV perception systems, leading to the development of diverse adversarial techniques [6, 7, 9, 12, 22, 41, 44, 48, 56–59]. Among them, adversarial object-based attacks [44, 55, 56, 58, 59] have garnered significant attention due to their practicality, flexibility, and stealth. These attacks manipulate AV perception by strategically placing common objects (also known as adversarial objects), such as cardboard structures, color patches, or metal foil, in the driving environment to alter the object detection model's output. For instance, adversarial color patches can deceive camera-based object detectors, while reflective materials can manipulate LiDAR-based perception by generating misleading point cloud data. The effectiveness of these attacks has been demonstrated against camera-based [9, 45], LiDAR-based [44, 55, 56, 59], radar-based [57], and sensor fusion systems [5, 6, 58] in real-world experiments, highlighting severe security concerns for AVs.

To mitigate the safety risks posed by these attacks, researchers have developed various defense mechanisms, primarily targeting

* Chenglin Miao is the corresponding author.



camera-only and LiDAR-only detection systems. For example, adversarial training [14, 28, 43] has been proposed to enhance the robustness of camera-based object detection by incorporating adversarial examples into the training dataset. However, this approach suffers from limited generalization and has been shown to be ineffective for other sensing modalities [53]. Similarly, while specific defenses have been proposed to counter LiDAR attacks [36, 53, 54], they are not generalizable to other sensing modalities or multi-sensor fusion systems. To the best of our knowledge, there have been no studies so far on how to address adversarial object-based attacks across different sensing systems, leaving a critical security gap in AV perception defenses.

This gap raises a crucial research question: **Is there a generalized defense capable of mitigating adversarial object-based attacks across different sensing modalities and their fusion?**

To answer this question, we conduct a comprehensive analysis of existing adversarial object-based attacks, aiming to identify their weaknesses and potential limitations that could inform the design of a generalized defense mechanism. Our analysis reveals that adversarial object-based attacks are not universally effective. While they succeed in most cases, there exist specific positions and regions along the road segment where the attack fails to achieve its intended impact. This observation is primarily attributed to two key factors. First, the sensory data of an AV can vary significantly depending on its movement and viewing angles. These variations affect how objects are detected, leading to inconsistencies in attack effectiveness. Second, attackers typically lack access to real-time sensor data from the victim AV. Instead, they approximate the AV's possible driving directions and use their own sensors to collect surrogate data offline, which is then used to optimize the properties (e.g., number, location, or color pattern) of adversarial objects. However, it is impractical for an attacker to account for all possible driving trajectories and positions the victim AV may take. As a result, adversarial objects are optimized for specific viewpoints, making them most effective when the AV follows the expected trajectories. If the AV deviates from these trajectories, the attack's success rate may decrease due to inconsistencies in perception.

Motivated by these insights, in this paper, we propose a novel defense mechanism that **strategically leverages motion planning** to counter adversarial object-based attacks. Instead of relying solely on perception-level defenses, our approach **integrates real-time attack detection with trajectory adaptation**, guiding the AV to positions where the attack is less effective.

The proposed defense consists of two key components. First, a deep reinforcement learning (DRL)-based motion planning model dynamically analyzes real-time perception outcomes along with the AV's motion status, such as speed and heading angle, and adjusts its trajectory to maximize attack failure. Second, an uncertainty-aware filtering scheme refines detection results by identifying and filtering unreliable perception outputs, enhancing the robustness of attack detection. To the best of our knowledge, **this is the first study to propose a generalized defense mechanism against different types of adversarial object-based attacks in autonomous driving**. Extensive experiments in both simulated and real-world environments demonstrate that our defense effectively detects object-based attacks across different sensor modalities and sensor fusion, while maintaining safe and comfortable driving behavior.

2 Background and Related Work

2.1 Autonomous Driving Perception

The autonomous driving pipeline typically consists of perception, planning, and control modules [5, 10, 15, 42, 51]. As a core function of the perception module, object detection is critical for ensuring system safety and functionality. Modern AVs use camera, LiDAR, and radar sensors for this task. Cameras capture RGB images that DNNs convert into bounding boxes, while LiDAR emits laser pulses to generate point clouds for predicting 3D bounding boxes [24, 39, 50]. Radar sensors transmit Frequency Modulated Continuous Wave (FMCW) signals and use Fourier processing to estimate object distance and angle [17, 18, 30, 32]. Object detection in autonomous driving may rely on a single modality or on multi-sensor fusion for improved accuracy and robustness [2, 6, 20, 27, 31, 33, 34, 46].

Existing camera- and LiDAR-based object detection systems generate bounding boxes with confidence scores, and an object is considered detected only when the score exceeds a predefined threshold. For radar-based detection, a common approach is to analyze the peak energy of the reflected echo signal and determine the presence of an object based on whether this peak exceeds a threshold [58]. Additionally, to reduce false alarms, these systems often adopt a strategy in which an object is considered detected only if it is perceived a certain number of times within a short time window. This strategy has been widely adopted in autonomous driving platforms such as Baidu Apollo and Autoware [1, 2].

2.2 Adversarial Object-Based Attacks

As described in Section 1, adversarial object-based attacks mislead the perception system by strategically placing common objects in the driving environment. The properties of these objects (e.g., number and location) are typically determined offline, after which the adversarial objects are deployed during the attack. Existing object-based attacks in autonomous driving can be broadly categorized into *single-sensor attacks* and *multi-sensor fusion attacks*.

Single-sensor attacks. Single-sensor attacks target individual perception modalities, such as cameras, LiDAR, or radar, by placing adversarial objects in the environment to mislead object detection models. Camera-based attacks [9, 12, 47, 48] typically use color patches. By carefully designing and positioning these patches, attackers can alter the visual appearance of targets, making them undetectable by the perception model. LiDAR-based attacks [44, 55, 56, 59] introduce adversarial objects with reflective surfaces (e.g., drones, road signs, or cardboard) to generate additional LiDAR points, deceiving the deep learning model used by the perception system. Radar-based attacks [57] manipulate reflected signals by placing metallic surfaces at specific orientations between the radar and a target object, resulting in incorrect detections.

Multi-sensor fusion attacks. Multi-sensor fusion systems are often considered a natural defense against single-sensor attacks. However, recent studies [6, 9, 58] show that a strong attack on a single modality can significantly compromise fusion system reliability [9]. Moreover, multi-sensor systems remain vulnerable to coordinated attacks targeting multiple modalities simultaneously. For instance, AdvFusion [58] integrates optimized color patches, metal foil, and cardboard into a single attack board that simultaneously deceives camera, LiDAR, and radar perception models. These

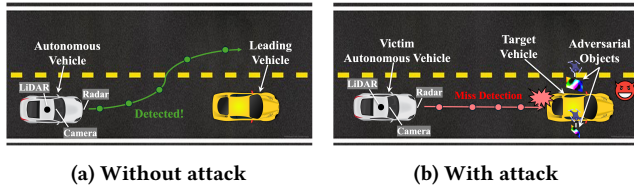


Figure 1: Vehicle hiding attack.

findings show that multi-sensor fusion alone is insufficient as a defense, highlighting the need for stronger countermeasures.

2.3 Defense Mechanisms

To mitigate adversarial object-based attacks in autonomous driving, various defenses have been proposed, primarily targeting single-sensor systems such as camera- or LiDAR-based detection. Adversarial training [14, 43] is widely used to improve robustness by incorporating adversarial examples into training, but it often reduces accuracy on benign inputs and cannot cover all possible perturbations. This challenge is more pronounced for LiDAR-based attacks, where the number and placement of adversarial objects vary across scenarios, making comprehensive training difficult. Although a few recent defenses [53, 54] target LiDAR attacks, they do not generalize to other modalities or fusion settings. To the best of our knowledge, no existing defense generalizes across different sensor modalities and fusion systems. In this paper, we address this gap by proposing a defense mechanism applicable to diverse sensor configurations in autonomous driving.

3 Problem Setting

Threat model. While many studies have explored adversarial object-based attacks on autonomous driving perception, most focus on the goal of *vehicle hiding* [7, 8, 41, 57–59]. As illustrated in Figure 1, a vehicle hiding attack aims to make a parked target vehicle disappear from the victim AV’s object detection system by strategically placing adversarial objects around it, potentially causing traffic collisions. To the best of our knowledge, only one recent study [26] has explored the vehicle creating attack on LiDAR perception, where the attacker attempts to generate a non-existent vehicle by placing objects on the roadside and using specialized equipment to interfere with LiDAR signals. Following most prior work, this paper focuses on the vehicle hiding attack as a representative scenario to illustrate our defense mechanism. The extension of our defense to the vehicle creating attack is discussed in Section 7.5.

For the *attacker’s capability*, we follow the settings of existing attacks and assume that the attacker does not have real-time access to the victim AV’s sensory data. To determine how to deploy adversarial objects, the attacker can collect surrogate sensory data offline by mimicking the victim AV’s potential driving behaviors and capturing data from different directions and distances relative to the target vehicle. Additionally, the attacker may have either white-box or black-box access to the object detection model.

For the *defender’s capability*, we consider a practical but challenging setting where the defender has no prior knowledge of the

adversarial objects, including their number, shapes, sizes, or locations. Furthermore, the defender does not know the road segment where the attack will occur. However, the defender, typically the manufacturer, is aware of the sensor modality or modalities used in the victim AV’s perception system.

Defense goal. Our objective is to develop a generalized defense mechanism capable of countering adversarial object-based attacks on both single-sensor and multi-sensor fusion systems. The mechanism should integrate into existing autonomous driving systems without requiring significant modifications to system modules.

4 Defense Challenges

We need to address several challenges to achieve the above goal.

Unknown adversarial objects. Adversarial object-based attacks vary in strategy, environment, and adversarial object properties, making them highly flexible. Since the defender lacks prior knowledge of these objects, designing an attack-agnostic defense that effectively handles unknown attack conditions is challenging.

Compatibility with various autonomous driving systems. AV systems differ in sensor modalities, detection algorithms, and planning mechanisms, making a universal defense challenging. The proposed mechanism should be adaptable and integrate seamlessly across different platforms without major system modifications.

Being efficient. Another challenge lies in the limited distance between the victim AV and the target vehicle available for defense. Due to the constrained sensing range of perception sensors, the AV may recognize the attack only when it is close to the target. Thus, the defense mechanism must operate efficiently, enabling accurate attack assessment and response within a short distance.

Minimizing disruptions to normal driving. Since attacks occur only occasionally, the defense mechanism should not operate continuously. In most cases, no attack is present, so the defense must minimize disruptions to normal decision making. An overly aggressive defense may trigger unnecessary interventions. However, determining when to activate the defense is challenging because the defender lacks knowledge of the attack timing and location.

5 Attack Analysis

Before presenting our methodology to address the above challenges, we first analyze existing attack methods to extract key insights that can guide the design of an effective defense mechanism. Specifically, in this section, we seek to answer the following key questions: (1) *What is the difference in detection performance with and without the attack?* (2) *In adversarial scenarios, what are the key factors that affect the attack performance?* (3) *Can existing adversarial object-based attacks achieve zero failure in all sensing conditions?* These questions are crucial for understanding the common patterns demonstrated in the attacks and developing effective defense mechanisms.

To investigate these questions, we conduct a case study on two representative vehicle hiding attacks: AdvLoc [59], which targets LiDAR object detection, and AdvFusion [58], which targets multi-sensor fusion. For AdvFusion, we use LiDAR-camera fusion as an example. Details of both methods are provided in Section 7. Following their original settings, we consider multiple possible driving directions of the victim AV when determining the placement and properties of adversarial objects. We implement these attacks with

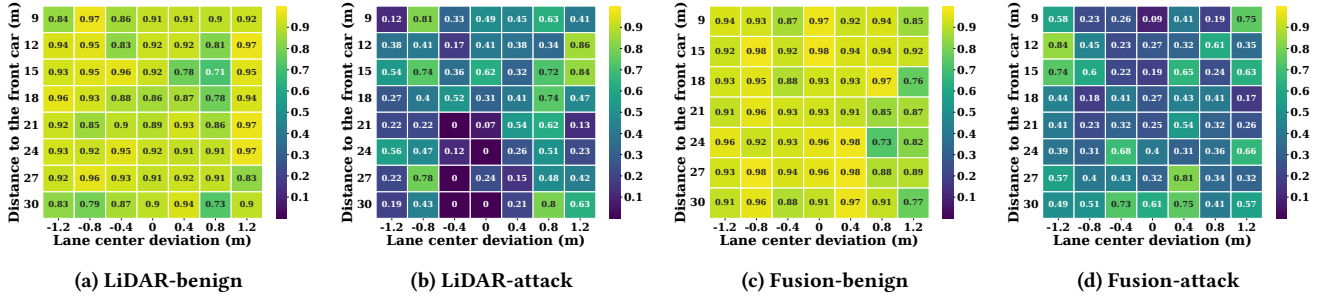


Figure 2: Heatmaps of average detection confidence in real-world scenarios.

a perception testbed equipped with a Velodyne VLP-32C LiDAR and a ZED 2i camera, as described in Section 8. Figures 11a and 11d illustrate these attacks in real-world scenarios. As discussed in Section 2.1, an object is detected in a frame only if its confidence score exceeds a predefined threshold (e.g., 0.5). The attacker’s goal is to reduce the confidence score below this threshold. To evaluate attack performance, we examine how confidence changes as the AV moves along the road, with an initial distance of 30 meters.

Figure 2 presents average confidence heatmaps. Figures 2a and 2c show the heatmaps for LiDAR-only and LiDAR-camera fusion detection systems without the attack (denoted as LiDAR-benign and Fusion-benign), while Figures 2b and 2d illustrate the corresponding maps under attack (denoted as LiDAR-attack and Fusion-attack). The horizontal axis represents the AV’s deviation from the lane center, and the vertical axis indicates its distance to the target vehicle. Each grid reflects the average detection confidence at a given region, with darker colors indicating lower confidence scores.

The results show that without an attack, both the LiDAR-only and LiDAR-camera fusion systems maintain stable, high confidence across most of the road segment. Under attack, however, confidence drops significantly in certain regions, with higher variance across locations. Although modeling confidence distributions is challenging, we observe key patterns. First, attack effectiveness varies by region and perspective. Neither detection system is consistently deceived across all areas, and even slight perspective changes can impact the attack performance. Second, while attacks reduce confidence, driving it to zero is difficult. Although those confidence scores may be below the detection threshold, they still carry meaningful information, reflecting attack impact variations over time and location. Third, confidence scores are not randomly distributed but tend to cluster into high or low-confidence regions.

We outline the primary reasons for the patterns we observed. First, sensory data varies significantly based on the AV’s movement and viewing angles, affecting object perception and reducing attack consistency. For instance, RGB images vary with perspective, distance, and lighting, causing discrepancies between optimized and actual captured values in real-world scenarios. Similarly, LiDAR point clouds are highly sensitive to viewpoint shifts. Unlike 2D images, 3D point clouds contain more depth information. Even slight changes in perspective can significantly alter the perceived surface, shape, and structure of an object, particularly for points affected by occlusion. Additionally, as the victim vehicle moves closer, the spacing between adjacent LiDAR points decreases and density increases,

which may impact attack effectiveness. Similar effects occur with radar because the AV’s position directly influences the reflected signal strength and direction. Second, it is nearly impossible to account for all possible driving trajectories or positions within a target road segment. As a result, adversarial objects are optimized for specific viewpoints and are most effective when the victim AV follows expected paths. Significant deviations from these paths can reduce attack success. In addition, even if attackers consider a broad range of driving directions, real-world environmental complexities introduce uncertainties, making it difficult to ensure consistent attack effectiveness across all viewing angles and distances. We evaluate the impact of the attacker’s efforts in Section 7.

In summary, the above analysis shows that while the attack is effective in most positions along the road, there are specific regions where it fails. As a result, the target vehicle may still be detected sporadically or consistently in certain areas. This presents an opportunity for defense. **If the victim AV can detect inconsistencies in confidence scores and be guided toward these specific regions, it may enable attack detection and facilitate countermeasures.** Unfortunately, if the target is detected only sporadically in a few frames, it may be dismissed as a false alarm. To suppress false detections, AV detection algorithms typically require an object to be detected multiple times within a short time window. Thus, sporadic detections alone are insufficient. Instead, the AV should be guided toward regions where the target can be consistently detected enough times within a short window, ensuring a valid object detection and enabling effective attack recognition.

However, a key challenge for this strategy is that the victim AV cannot determine the detection confidence distribution for a road segment before traversing it. Additionally, the adversarial objects are unknown in advance, making it impossible to pre-identify regions where the attack will fail. To address this, an **intelligent defense mechanism** is needed to dynamically adjust the AV’s driving behavior by continuously analyzing real-time detection confidence variations and its motion status (e.g., speed, acceleration, and heading angle), allowing it to **navigate toward regions where the attack is less effective.**

6 Methodology

6.1 Overview

Building on the above analysis, we propose a novel defense mechanism that leverages **intelligent motion planning** to counter the

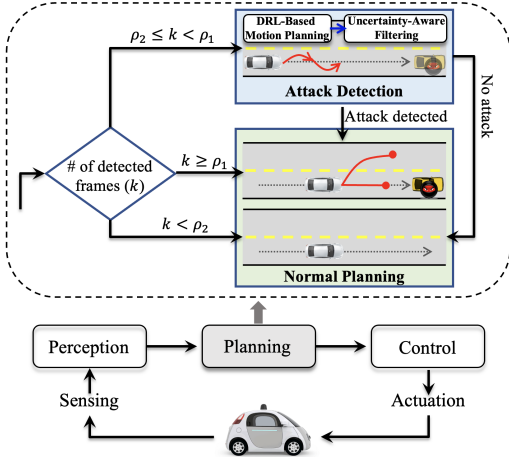


Figure 3: Defense mechanism overview.

attack. Figure 3 provides an overview of the mechanism, which is integrated into the AV's planning module. At its core is the **attack detection component**, consisting of a *DRL-based motion planning model* and an *uncertainty-aware filtering scheme*. The DRL-based motion planning model analyzes real-time perception outcomes and the AV's motion status to dynamically adjust its trajectory, guiding it toward regions where the attack is less effective. When motion planning alone fails to detect an attack, the uncertainty-aware filtering scheme reanalyzes historical perception data, filtering unreliable outputs to enhance detection robustness. To ensure broad applicability, the **normal planning component** represents the AV's built-in planning model, treated as a black box. This design choice ensures that the defense mechanism remains generalizable and seamlessly integrates with existing AV planning systems.

When the victim AV operates on the road, the planning module continuously processes perception results. A key challenge is determining when to activate the attack detection component. As noted earlier, adversarial object-based attacks are imperfect, and the target vehicle may occasionally appear in some frames. While usually dismissed as false alarms, these detections offer an opportunity for attack detection. To formalize this, we define an object detection threshold ρ_1 , representing the minimum number of detected frames within a time window required for a valid object detection. Here, the time window is defined as T consecutive frames. To capture sporadic yet suspicious object detections, we introduce a defense activation threshold ρ_2 ($\rho_2 < \rho_1$) as a trigger for activating the attack detection component. As shown in Figure 3, if the number of frames where the target vehicle is detected within T frames, denoted as k , satisfies $\rho_2 \leq k < \rho_1$, the attack detection model is activated. Otherwise, detections in these frames are classified as false alarms ($k < \rho_2$) or valid object detections ($k \geq \rho_1$), allowing the AV to proceed with its normal response.

Once the attack detection component is activated, the victim AV reduces its speed to a safe level. The DRL-based motion planning model then analyzes real-time perception outcomes/motion dynamics and adjust the AV's trajectory to increase the likelihood of attack failure while ensuring safe and comfortable driving. After operating

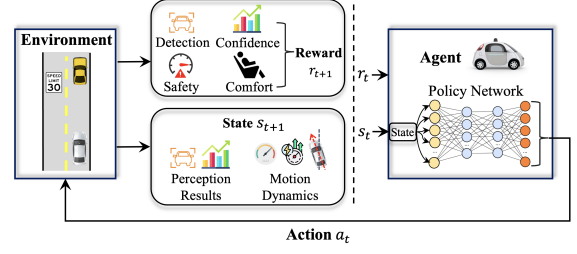


Figure 4: DRL-based motion planning model.

for a period, the system determines whether an attack is present. Specifically, if the target vehicle is detected in more than ρ_1 frames within T consecutive frames before the AV reaches a predefined safe distance, which allows sufficient time for evasive actions like lane changes or braking, the attack is confirmed. This safe distance is estimated based on the sporadic detections observed before activating attack detection. If the attack is not confirmed upon reaching this distance, the uncertainty-aware filtering scheme is triggered to further analyze perception data and verify the presence of an attack. As shown in Figure 3, if an attack is finally detected, the normal planning model guides the AV to evade it (e.g., by changing lanes). Otherwise, detections are classified as false alarms, and the AV proceeds without evasive maneuvers.

6.2 DRL-Based Motion Planning

A key challenge in designing the motion planning model is that the victim AV has no prior knowledge of the number, locations, or shapes of adversarial objects, nor can it identify regions where attacks are less effective. Without this information, how can the AV adjust its motion to evade the attack? To address this, we propose a motion planning model driven by real-time environmental feedback. Specifically, the AV uses perception outputs and their confidence levels to dynamically guide its movement. To implement this model, we adopt DRL to generate an optimal motion policy. DRL enables an agent to interact with an environment and learn a policy that maximizes rewards, making it well-suited for autonomous driving tasks. Figure 4 illustrates our DRL-based motion planning model. At each time step t , the agent (i.e., the victim AV) observes a state s_t , which includes perception results and motion dynamics (e.g., speed, acceleration, and heading angle). This state is fed into a policy network that outputs an action a_t . The agent executes this action, transitions to a new state s_{t+1} , and receives a reward r_{t+1} reflecting environmental feedback. The reward function accounts for perception outcomes, driving safety, and passenger comfort. The agent's objective is to learn a policy that maximizes the expected cumulative reward over time. Using this trained policy, the AV can take a series of actions to both maximize attack detection and maintain safe, comfortable driving. Below we elaborate on the state space, action space, and reward function.

6.2.1 State Space. The agent's state at time step t is represented as a vector: $s_t = (k_t, p_t, c_t, \Delta c_t, d_t, v_t, \phi_t, \psi_t)$. Here, k_t denotes the number of frames in which the target vehicle is detected within the past T frames. The perception result p_t is set to 1 if the vehicle is detected ($k_t \geq \rho_1$) and 0 otherwise. The detection confidence

score at time step t is denoted as c_t , with its change from the previous step given by $\Delta c_t = c_t - c_{t-1}$. The AV's lateral offset from the lane center, d_t , is computed using its center of mass and road waypoints (e.g., derived from HD maps and GPS-IMU fusion). The state also includes the AV's speed v_t , acceleration ϕ_t , and heading angle deviation ψ_t relative to the road direction.

6.2.2 Action Space. The agent's action space consists of steering and longitudinal control, enabling precise motion regulation. At time step t , the action is represented as $a_t = (\delta_t, \tau_t, \beta_t)$, where δ_t is the steering intensity, τ_t is the throttle intensity, and β_t is the braking intensity. This formulation aligns with modern autonomous driving systems, allowing direct control over vehicle dynamics. Given the current state, the DRL agent selects an action using a policy network, which models a probability distribution $\pi(a_t|s_t, \theta)$ over the action space. We implement this policy network as a DNN with learnable parameters θ . It predicts a Gaussian distribution for each action dimension. Specifically, the steering action is sampled as $\delta_t \sim \mathcal{N}(\mu_\delta(s_t, \theta), \sigma_\delta(s_t, \theta))$. For longitudinal control, the network outputs a single Gaussian distribution, from which the longitudinal action Λ_t is sampled as $\Lambda_t \sim \mathcal{N}(\mu_\Lambda(s_t, \theta), \sigma_\Lambda(s_t, \theta))$. Here $\mu_\delta(s_t, \theta)$ and $\mu_\Lambda(s_t, \theta)$ represent the predicted mean values for steering and longitudinal control, respectively, while $\sigma_\delta(s_t, \theta)$ and $\sigma_\Lambda(s_t, \theta)$ denote their corresponding standard deviations. The throttle and brake intensities are determined based on the sampled longitudinal control action Λ_t . If Λ_t is positive, it is assigned to the throttle τ_t , while the brake β_t remains zero. Conversely, if Λ_t is negative, its absolute value is assigned to the brake β_t , and the throttle τ_t is set to zero. This formulation ensures that throttle and braking are mutually exclusive, preventing simultaneous activation.

6.2.3 Reward Function. During training, the agent receives a reward based on its state after taking an action. The reward function is designed to align with the system's objective. Since the motion planning model aims to maximize attack exposure while ensuring driving safety and passenger comfort, the reward function consists of the following components.

Detection reward. This reward evaluates whether an action enhances attack detection. Since successful detection is the cumulative result of multiple detections within a time window, we include both the target detection result p_t and detection confidence c_t into the reward. To further encourage the model to capture positive changes, we introduce an additional incentive when detection confidence increases significantly. The detection reward is defined as:

$$r_t^d = \mu_1(\gamma p_t) + \mu_2 c_t + \mu_3 \varphi(\Delta c_t), \quad (1)$$

where $\{\mu_i\}_{i=1}^3$ are weighting parameters balancing the three terms. A discount factor γ is applied to p_t , ensuring that later detections yield progressively lower rewards. This discount factor serves two purposes. First, it prevents the agent from exploiting static positions (e.g., braking and staying idle) to accumulate excessive rewards, a common issue in reinforcement learning where agents exploit risk-averse strategies. Second, since later detections often rely on previous frames due to overlapping time windows, rewarding earlier detections encourages proactive exploration and adaptive driving, preventing overly cautious or stationary behaviors. The function $\varphi(\cdot)$ quantifies whether the AV transitions from a low- to a high-confidence region, ensuring only significant confidence increases

contribute to the reward. If the confidence change Δc_t exceeds a predefined threshold h , $\varphi(\Delta c_t)$ equals Δc_t . Otherwise, it is set to 0, preventing minor fluctuations from influencing the learning.

Safety and comfort reward. Beyond mitigating attacks, our defense mechanism prioritizes vehicle safety and driving comfort during the defense process. To enforce these objectives, we define a reward function that penalizes unsafe behaviors while promoting smooth and efficient driving:

$$\begin{aligned} r_t^s = & -\omega_1 \mathbb{I}(d_t \notin [0, d_{\max}]) - \omega_2 \mathbb{I}(v_t \notin [0, v_{\max}]) \\ & - \omega_3 \mathbb{I}(\Delta \phi_t \notin [0, \Delta \phi_{\max}]) - \omega_4 \delta_t'^2 \\ & - \omega_5 |\delta_t'| v_t^2 + \omega_6 v_t, \end{aligned} \quad (2)$$

where $\{\omega_i\}_{i=1}^6$ are weighting parameters that balance the contributions of different terms. The indicator function $\mathbb{I}(d_t \notin [0, d_{\max}])$ is defined as

$$\mathbb{I}(d_t \notin [0, d_{\max}]) = \begin{cases} 1, & \text{if } d_t \notin [0, d_{\max}], \\ 0, & \text{if } d_t \in [0, d_{\max}], \end{cases}$$

where $d_{\max}$ represents a safe distance that ensures the AV remains within the same lane. The definitions of $\mathbb{I}(v_t \notin [0, v_{\max}])$ and $\mathbb{I}(\Delta \phi_t \notin [0, \Delta \phi_{\max}])$ follow a similar structure. $v_{\max}$ denotes the predefined speed warning for the DRL model, applying a penalty when it is above this value. After the defense is activated, The AV first checks whether its current speed exceeds $v_{\max}$ and reduces it to $v_{\max}$ if necessary. $\Delta \phi_t = |\phi_t - \phi_{t-1}|$ represents the change in acceleration, with $\Delta \phi_{\max}$ as its predefined upper bound. δ_t' represents the clipped steering intensity, ensuring it remains within a safe operational range. This reward function comprises six terms to enforce driving safety and comfort. The first three penalize lane deviations, excessive speeds, and abrupt acceleration changes. The fourth and fifth terms discourage large steering angles, with additional penalties for sharp turns at high speeds to maintain stability. The final term incentivizes movement, preventing the AV from lingering at low speeds, which could pose safety risks. This formulation balances safety, comfort, and mobility, ensuring stable and realistic autonomous driving behavior.

The final reward of the agent at time step t is calculated as $r_t = r_t^d + r_t^s$. To train the policy network, we simulate existing adversarial object-based attacks across various driving environments. For the DRL model, we employ Soft Actor-Critic (SAC) [19], a state-of-the-art stochastic off-policy method, to optimize the policy in a continuous action space.

6.3 Uncertainty-Aware Filtering

While the DRL-based motion planning model effectively guides the AV to counteract attacks, challenging scenarios may still lead to missed attack detections. To address this, we introduce an uncertainty-aware filtering scheme that enhances defense capability. Specifically, if no attack is detected after the AV reaches a predefined safe distance, the system performs an additional verification step by filtering unreliable object detection results along the defense trajectory. Our investigation reveals that attack effects are often localized, causing low object detection confidence in certain regions while neighboring areas maintain higher scores. In contrast, confidence values in benign environments typically remain smooth and consistent across neighboring regions. If the AV traverses such

adversarially affected regions during the defense process, the resulting low-confidence detections may cause missed target vehicle detections and degrade defense performance.

To improve reliability, we propose filtering out unreliable detections from these regions. However, identifying them is challenging, as it requires comparing confidence scores with neighboring regions. Since the AV can only observe detection confidence at its current locations and lacks access to unexplored neighboring regions, direct comparison is not feasible. To overcome this, we divide the trajectory into N segments $\{\mathcal{S}_j\}_{j=1}^N$ and train a Long Short-Term Memory (LSTM) model [21] to predict the most unreliable segment based on the observed confidence scores and locations. For training, we implement existing adversarial object-based attacks and generate object detection confidence heatmaps for each attack scenario, as in Figure 2, with grid sizes of $0.4\text{ m} \times 3\text{ m}$. The unreliability score for each grid is computed as $u_g = \frac{1}{|\mathcal{B}_g|} \sum_{b \in \mathcal{B}_g} (\bar{c}_b - \bar{c}_g)$, where $\bar{c}_g$ is the average detection confidence in grid g , $\mathcal{B}_g$ represents its neighboring grids, and $\bar{c}_b$ is the average confidence score in neighboring grid b . We then simulate multiple defense trajectories and assign ground-truth labels to the most unreliable segments based on the average unreliability scores of the grids each segment traverses. Additionally, to account for benign conditions, we include scenarios without attacks and assign the ground-truth label as “None”.

In practice, model predictions are often uncertain due to factors like attack strength and environmental variations. While the classification model identifies the most unreliable segment, its predictions may be imprecise or overly confident. To improve prediction reliability, we incorporate conformal prediction [38] into our framework to provide principled uncertainty quantification and set-valued predictions. At a significance level α , conformal prediction guarantees marginal coverage of at least $1 - \alpha$, meaning that the prediction set contains the true segment with probability at least $1 - \alpha$ under the data distribution. To apply conformal prediction, we reserve a portion of the training data as a calibration set. For each calibration sample (x_i, Y_i) , we compute a nonconformity score $\epsilon_i = 1 - P_{Y_i}(x_i)$, where $P_{Y_i}(x_i)$ denotes the predicted probability assigned to the true segment Y_i . Larger values of ϵ_i indicate lower conformity and higher uncertainty. Given a predefined significance level α , we compute a threshold η_α from the calibration nonconformity scores using the standard conformal quantile rule.

During testing, for a new defense trajectory x and each candidate segment $\mathcal{S}_j$, we compute the nonconformity score $\epsilon_j = 1 - P_{\mathcal{S}_j}(x)$, and construct the conformal prediction set $\mathcal{S}_{\text{pred}} = \{\mathcal{S}_j \mid \epsilon_j \leq \eta_\alpha\}$. When the model is uncertain, the set tends to include multiple segments. When it is confident, the set is smaller. We then filter out frames within the predicted unreliable segments in $\mathcal{S}_{\text{pred}}$ and determine whether an attack is present based on the remaining frames. Specifically, let N_r denote the number of remaining frames, among which the target vehicle is detected in K frames. An attack is confirmed if $K/N_r \geq \rho_1/T$; otherwise, the attack is not detected.

7 Performance Evaluation

We begin by evaluating our defense mechanism using CARLA [11], a widely used open-source simulation platform for developing and testing autonomous driving algorithms.

7.1 Experimental Setup

Sensor setup. We configure the CARLA LiDAR sensor with 64 channels, a 26.8° vertical field of view, a 10 Hz frequency, and a 100 m range, generating approximately 1.3 million points per frame. Both the LiDAR and camera sensors are mounted on top of the AV, with the LiDAR sensor positioned at 1.8 m above the ground. The camera sensor, capturing 1024×1024 resolution images with a 110° field of view, is mounted at a height of 1.68 m . We simulate the radar signal with mesh-based methods [23, 58], and the signal is further processed into range-azimuth map (RAMAP) with a resolution of 100×80 . The height of the Radar sensor is set to 1 m .

Driving environment setup. We consider diverse driving scenarios in CARLA, including both urban and rural environments with varying infrastructure and road conditions. We select 80 scenes in total, with 50 for training and 30 for testing. The target vehicle is parked $40\text{--}60\text{ m}$ ahead of the victim AV on a common two-lane road, where each lane is approximately 3.7 m wide.

Perception models and attack methods. We adopt PointPillar [24] and PIXOR [50] for LiDAR-based object detection, and YOLOv3 [35] and Fast R-CNN [13] for camera-based detection. For radar, we use the Constant False Alarm Rate (CFAR) algorithm [37]. For sensor fusion, we employ Weighted Box Fusion (WBF) [40], a decision-level fusion mechanism. We consider four state-of-the-art object-based attacks: AdvLoc [59], AE-Morpher [56], and BALiDAR [55] as LiDAR-only attacks, and AdvFusion [58] as a fusion-based attack. AdvLoc identifies adversarial locations where placing common objects disrupts LiDAR perception. AE-Morpher extends earlier object-based attacks [44] by modifying object geometry to improve real-world effectiveness. BALiDAR differs from these attacks by compromising both training and testing [16, 25, 52] through backdoor point clouds, where objects placed on the vehicle roof act as triggers to manipulate perception. For sensor fusion, AdvFusion combines adversarial patches, metal foil, and cardboard into a single attack board to attack LiDAR, camera, and radar simultaneously. When deriving parameters (e.g., number and location) of adversarial objects, we consider three possible driving directions of the victim AV: one along the lane center and two randomly selected directions on the left and right of the lane center.

Baseline methods. To the best of our knowledge, no existing defense method is applicable across different sensor modalities or fusion. We adopt *adversarial training*, *random trajectory*, and *weaving trajectory* as baseline defenses. In adversarial training, we retrain the object detection model using adversarial data generated by simulating existing object-based attacks. In the random trajectory baseline, the victim AV follows a randomized trajectory within the same road segment and starting point used to evaluate our mechanism. In the weaving trajectory baseline, the victim AV moves side to side in an S-shape within the road segment.

Evaluation metrics. We primarily use the following metrics to evaluate defense performance:

- **Detection rate (DR).** This metric is the percentage of defense trials in which the attack is detected across all trials conducted under various attack scenarios. A trial begins when the defense is activated and ends with the determination of whether an attack is present. A higher DR indicates a more effective defense.

Table 1: Definition of discomfort driving incidents.

Incident	Definition
Overspeeding	$v_t \geq v_{max} + 10 \text{ km/h}$
Low-speed driving	$v_t \leq \frac{v_{max}}{2}$
Large lane deviation	$d_t \geq 1.5 \text{ m}$
Large heading deviation	$\psi_t \geq 15^\circ$
Abrupt acceleration change	$ \phi_t - \phi_{t-1} \geq 0.5 \text{ m/s}^2$

- *Remaining distance (RD)*. This metric is the distance between the victim and the target vehicle when the attack is detected. A longer RD indicates more space to react and take countermeasures.
- *Discomfort driving incident rate (DDIR)*. This metric represents the percentage of defense trials in which at least one discomfort incident occurs. As shown in Table 1, these incidents include overspeeding, low-speed driving, large lane deviation, and abrupt heading angle or acceleration changes [4, 29].

Parameter settings. We set the time window length T to 10, the object detection threshold ρ_1 to 5, the defense activation threshold ρ_2 to 2, and the object detection model confidence threshold to 0.5. In the reward function, we set the values of $\{\mu_i\}_{i=1}^3$ as 80, 2, and 5, respectively, for the detection reward described in Section 6.2. The confidence increase threshold h is set to 0.3. The discount factor γ is set to $1/n$, where n represents the number of times the target vehicle has been detected up to time step t . For the safety and comfort reward, the values of $\{\omega_i\}_{i=1}^6$ are set to 5, 10, 5, 3, 0.2, and 1, respectively. We set d_{max} , v_{max} , and $\Delta\phi_{max}$ as 1.5 m, 15–25 km/h according to evaluation, and 0.5 m/s^2 , respectively. For uncertainty-aware filtering, the detection trajectory is divided into 10 segments (i.e., $N = 10$), and the significance level α is set to 0.15. The predefined safe distance is 8 m, allowing lane changes or emergency braking at speeds of 30 km/h or lower [3]. Our DRL-based motion planning model is trained for 80 epochs on two NVIDIA GeForce RTX 6000 GPUs. The policy and Q-value learning rates are set to 0.001, and the entropy regularization coefficient is 0.1. Additionally, our algorithms, driving environments, and low-level agent controls interact in real-time via the user-friendly APIs provided by SafeBench [49].

7.2 Analysis of Defense Activation

As described in Section 6.1, sporadic detections indicate potential attacks and trigger our motion planning model. In our design, the model is activated when $\rho_2 \leq k < \rho_1$. If ρ_2 is too low (e.g., 1), temporary objects or misdetections (e.g., billboards identified as vehicles) may cause false positives and unnecessary defense activations. Conversely, a high ρ_2 risks failing to activate the defense. To determine an appropriate ρ_2 , we randomly select 80 scenes across four town maps in CARLA using PointPillar for LiDAR detection. For each scene, we evaluate attack and benign scenarios and implement the AdvLoc attack in all attack scenarios. For the benign scenario, we divide the 80 scenes into two groups. In the first group (40 scenes), the environment remains unchanged. In the second group (40 scenes), we place a roadside billboard that may cause sporadic false detections of a front vehicle. The AV drives through these scenes and, based on our activation strategy, we compute the

Table 2: The number of missed defense activations in attack scenarios and unnecessary activations in benign scenarios.

ρ_2	1	2	3	4
# of missed defense activations (Attack)	0	0	3	9
# of unnecessary activations (Benign-Group1)	0	0	0	0
# of unnecessary activations (Benign-Group2)	5	2	1	0

Table 3: False alarm rates in benign and attack scenarios.

Distance range	Benign			Attack	
	No object	Billboard	Bus shelter	AdvLoc	AdvFusion
45–35 m	0	0.024	0.009	0.235	0.247
55–45 m	0	0.007	0.004	0.266	0.291

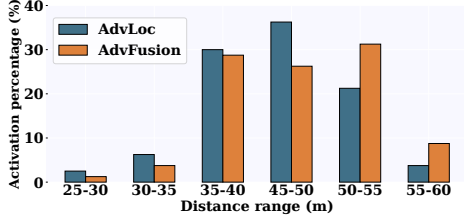
number of missed defense activations in attack scenarios and unnecessary activations in benign scenarios. Table 2 presents results for ρ_2 values from 1 to 4. As observed, a larger ρ_2 leads to more missed activations, while a smaller ρ_2 results in more unnecessary activations. Considering safety and driving smoothness, we set $\rho_2 = 2$ in this paper. The unnecessary activations also highlight the importance of the motion planning model in distinguishing actual threats from benign anomalies.

To further evaluate the designed activation strategy, we then quantify the false alarm rate in both benign and attack scenarios. For the benign scenario, we consider two cases: (1) no object in front of the AV, and (2) the presence of a roadside object (an advertising billboard or a bus shelter). For the attack scenario, we consider AdvLoc and AdvFusion attacks. In both scenarios, the false alarm rate is defined as the ratio of LiDAR frames in which a front vehicle is detected to the total number of LiDAR frames. We randomly select 80 scenes in CARLA and compute the average false alarm rate. Results are shown in Table 3. We consider different distance ranges between the AV and either the roadside object or the target vehicle. We can observe that benign roadside objects can occasionally be misclassified as vehicles, but the probability is low, with the false alarm rate below 0.025 when approaching such objects. This rate is insufficient to cause concentrated detections within a time window, resulting in only two unnecessary defense activations. In contrast, under attack the false alarm rate is about ten times higher than in benign scenarios, indicating that even strong attacks cannot fully suppress sporadic detections. With $\rho_2 = 2$, the defense is activated in 100% of the 80 scenes. These results demonstrate the effectiveness and practicality of the activation strategy.

Next, we analyze the typical locations of the victim AV when the motion planning model is activated. We set $\rho_2 = 2$ and implement AdvLoc and AdvFusion in each scene. Figure 5 shows the distribution of defense activations across different distances between the victim AV and the target vehicle. We can see that the defense is activated in all attack scenarios. Activation typically occurs at relatively long distances, usually within the 35–55 m range for both attacks. The figure shows that 86.7% of activations occur within this range, providing sufficient time and distance for defensive actions. Although activation sometimes occurs at 25–30 m, it still provides

Table 4: Defense performance on different object detection systems.

Initial distance	Speed (km/h)	LiDAR				LiDAR-camera fusion			
		DR with M(%)	DR with M+U(%)	RD(m)	DDIR(%)	DR with M(%)	DR with M+U(%)	RD(m)	DDIR(%)
40 m	15	83.0	93.2	12.9	0.4	76.3	85.8	14.0	0.2
	20	73.4	87.5	11.3	0.8	68.8	80.9	12.6	0.2
	25	70.7	83.6	11.0	0.3	66.3	78.5	12.4	0
50 m	15	87.1	95.5	14.8	0.3	80.4	87.7	15.9	0
	20	75.5	88.7	13.8	0.7	72.1	82.2	14.2	0.2
	25	71.6	83.4	13.3	0.4	70.2	79.6	13.5	0.1

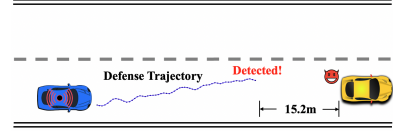
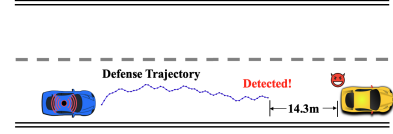
**Figure 5: The distribution of defense activations across different distance ranges between the victim AV and the target.**

enough time for the AV to respond. We evaluate the short-distance scenarios in Section 7.3.

7.3 Defense Performance on LiDAR Perception and Sensor Fusion

Overall performance. In this experiment, we evaluate the defense performance on LiDAR-only and LiDAR-camera fusion systems. We adopt PointPillar as the LiDAR detection model and YOLOv3 as the camera-based detection model, employing the WBF mechanism for fusion. For attack methods, we use AdvLoc for LiDAR-only detection and AdvFusion for LiDAR-camera fusion. In AdvLoc, we place up to three adversarial objects, each sized $0.45\text{ m} \times 0.45\text{ m} \times 0.2\text{ m}$. In AdvFusion, we use the same-sized adversarial objects for attacking LiDAR and add a color patch measuring $0.8\text{ m} \times 0.45\text{ m}$ to mislead the camera perception. For each detection system, we train our motion planning model in 50 scenes with the corresponding attack and evaluate it in 30 additional scenes. Each scene includes 30 attack trials with varying object placement. The results are presented in Table 4, where we evaluate two scenarios with initial distances of 40 m and 50 m between the victim AV and the target vehicle. For each scenario, we consider three cases where the victim AV’s speed at the start of attack detection is 15 km/h, 20 km/h, and 25 km/h. The table reports the average results over 30 evaluated scenes for each case. We present the DR obtained using only the motion planning model from Section 6.2 (denoted as ‘DR with M’) and the DR achieved by combining motion planning with uncertainty-aware filtering (denoted as ‘DR with M+U’).

The results demonstrate that our proposed attack detection model performs well across different object detection systems. For instance, using only the motion planning model, the DR reaches 87.1% for the LiDAR system and 80.4% for the LiDAR-camera fusion system when the initial distance is 50 m and the speed is 15

**Figure 6: Example of the defense trajectory (AdvLoc).****Figure 7: Example of the defense trajectory (AdvFusion).**

km/h, indicating that the model effectively navigates the victim AV to regions where the attack is more likely to fail. Furthermore, incorporating uncertainty-aware filtering further improves the DR, increasing it to 95.5% for LiDAR and 87.7% for LiDAR-camera fusion, respectively. This improvement suggests that the conformal prediction process successfully removes unreliable perception results. Moreover, the results indicate that the remaining distance (RD) in all cases is sufficient for the victim AV to take appropriate actions after the attack is detected. We also observe that the DR for LiDAR-camera fusion is slightly lower than that for LiDAR-only. This is likely because attacks targeting fusion systems consider observations from both LiDAR and camera sensors, resulting in a wider and more evenly distributed attack coverage compared to LiDAR-only attacks, thereby making the defense more challenging.

Table 4 further highlights the effectiveness of our mechanism in maintaining both safety and driving comfort. Specifically, the DDIR remains below 1% in all cases, indicating a low occurrence of discomfort-inducing driving behaviors. Additionally, we observe that lower vehicle speeds result in better overall defense performance, primarily because slower speeds provide more time for the mechanism to adjust the AV’s trajectory and gather sufficient observations to support both motion planning and conformal prediction processes. In Figure 6 and Figure 7, we visualize two examples of defense trajectories generated by our motion planning model for AdvLoc and AdvFusion. The victim AV’s starting points are randomly selected, with an initial distance of approximately 40 m from the target vehicle. The attack is detected at around 15 m. Note that the lane width is set to 3.7 m, and we provide a width-rescaled

Table 5: Comparison with baseline methods.

Defense method	LiDAR		LiDAR-camera fusion	
	DR(%)	RD(m)	DR(%)	RD(m)
Adversarial training	14.3	10.1	23.2	11.3
Random trajectory	31.5	10.5	20.9	10.9
Weaving trajectory	30.4	10.3	25.7	10.5
Ours	93.2	12.9	85.8	14.0

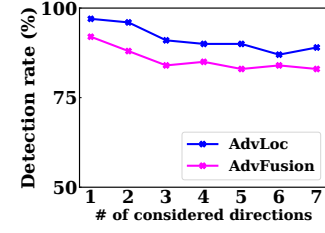
Table 6: Defense performance on different attack methods and object detection models.

Attack method	PointPillar		PIXOR	
	DR(%)	RD(m)	DR(%)	RD(m)
AdvLoc	93.3	13.8	91.9	15.7
AE-Morpher	87.2	12.2	79.1	12.9
BALiDAR	90.5	16.6	84.8	15.6
AdvFusion (w/ Fast R-CNN)	89.0	14.2	87.2	14.8

view of the trajectory for better visualization. We can see that the trajectories exhibit realistic motion with minor lateral adjustments.

Comparison with baseline methods. Next, we compare our defense with baseline methods under the same evaluation settings as Table 4. For adversarial training, we augment the object detection model’s training set with 10% adversarial data. Since attacks target these retrained models, we define DR for adversarial training as the average rate reduction in successful attacks when the AV moves forward from various directions. For the random trajectory, the AV remains within the same lane for safety but follows a random steering value, with corrective steering applied if it exceeds road boundaries. For the weaving trajectory, the AV moves side to side in an S-shape with a maximum heading angle deviation of 12° . All defenses are evaluated under identical attack settings with an initial speed of 15 km/h . Table 5 shows that adversarial training performs the worst (DR of 14.3% for LiDAR and 23.2% for LiDAR-camera fusion) due to its inability to cover all adversarial object variations. The random trajectory and weaving trajectory method offer some defense by causing AV deviations from optimal attack conditions but remains significantly less effective than ours.

Performance on different attack methods and object detection models. We evaluate our defense against three state-of-the-art LiDAR attacks: AdvLoc, AE-Morpher, and BALiDAR, using PointPillar and PIXOR as object detection models. For LiDAR-camera fusion, we replace YOLOv3 with Fast R-CNN as the camera detection model. Each attack follows its original implementation. BALiDAR uses a 0.8 m cube as a backdoor trigger, while AE-Morpher optimizes adversarial point positions and places adversarial objects at the rear or top of the target vehicle. For LiDAR-only detection, the victim AV is unaware of the attack method, so we train the motion planning model on all three attacks and test its defense performance against each. Table 6 presents the average DR after uncertainty-aware filtering across 30 test scenes. The distance between the AV and the target vehicle is 40 m , with a speed of 15 km/h post-defense

**Figure 8: Defense performance with varying numbers of attacker-considered driving directions for the victim AV.****Table 7: The DR(%) with/without uncertainty-aware filtering under different values of ρ_1 and the confidence threshold λ_m .**

	$\rho_1 = 4$	$\rho_1 = 5$	$\rho_1 = 6$
$\lambda_m = 0.4$	92/92	88/95	84/91
$\lambda_m = 0.5$	90/96	83/93	77/85
$\lambda_m = 0.6$	86/93	74/86	67/77

activation. Results confirm that our defense performs well across different attack methods and object detection models.

Impact of the number of potential driving directions considered by the attacker when designing adversarial objects. As described in Section 7.1, our experiments assume the attacker considers three potential driving directions for the victim AV. To assess how this number affects defense performance, we vary it from 1 to 7 and evaluate our mechanism against AdvLoc and AdvFusion attacks. These directions are randomly selected within the lane. To clearly illustrate the impact, we conduct this evaluation without applying uncertainty filtering. The initial distance between the two vehicles is set to 40 m , and each case involves 100 simulated attacks. Figure 8 presents the average attack detection rate. When the attacker optimizes for only 1–2 directions, the detection rate is noticeably higher than when multiple directions are considered. However, even when the attacker attempts to cover the entire lane (using more than 3 directions), our defense can still achieve good defense performance. For example, in AdvLoc, the DR is 91% for 3 directions and 87% for 6 directions. These results suggest that increasing the number of considered directions does not necessarily strengthen the attack. They also indicate that even if the attacker is aware of the defense mechanism and tries to cover more potential driving directions to maximize the attack effect, our defense can still be activated and remain effective.

Impact of the object detection threshold ρ_1 and the confidence threshold. We evaluate the impact of the valid object detection threshold, ρ_1 , and the object detection model’s confidence threshold (λ_m) on defense performance. Higher values for these parameters impose stricter conditions for recognizing the target vehicle. We conduct 100 experiments following the settings in Table 4 and report the average DRs before and after applying the uncertainty-aware filtering scheme. Table 7 presents the DRs under different parameter values. The results show that DR decreases as detection requirements become stricter, with lower ρ_1 and λ_m leading to higher DRs. However, even in stringent settings, our

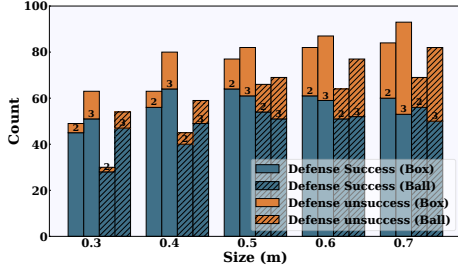


Figure 9: Attack and defense performance across different adversarial object shapes, numbers, and sizes.

Table 8: Defense performance at short distances.

Speed	LiDAR		LiDAR-camera fusion	
	DR(%)	RD(m)	DR(%)	RD(m)
15 km/h	86.2	10.8	81.5	10.5
20 km/h	81.5	9.6	77.7	9.3

defense maintains a DR above 85% in most cases after applying the uncertainty-aware filtering scheme, demonstrating its robustness.

Impact of variations in the shape, number, and size of adversarial objects. We further analyze the impact of adversarial object characteristics, including shape, size, and number, on defense performance. The number of objects are denoted in the figure (2 or 3). The experimental setup follows that of the previous section, with 100 experiments conducted for each setting, and we report the attack and defense performance under varying conditions. We evaluate both box- and ball-shaped adversarial objects, with lengths/diameters ranging from 0.3 m to 0.7 m. Figure 9 presents the attack and defense performance across different factors. We can see that our defense remains highly effective, achieving, for example, an 85.6% DR when using ball-shaped adversarial objects with a 0.4 m diameter. The results also show that box-shaped objects generally lead to a higher attack success rate and a lower DR than ball-shaped ones of similar size. While DR decreases as attack strength increases, our defense still maintains a 72% DR when three adversarial objects of 0.6 m in size are used.

Impact of activation distance. As shown in Figure 5, there are instances where the DRL-based motion planning model is activated at a distance of 25–30 m from the target vehicle. This experiment evaluates whether our defense remains effective at such short distances. Specifically, we follow the experimental setup in Table 4 and set the initial distance between the two vehicles to approximately 30 m. Table 8 reports the average DRs and RDs when the AV’s speed at defense activation varies from 15 km/h to 20 km/h. The results show that our defense remains effective even at shorter distances, as the victim AV still has enough space to execute countermeasures. This is largely due to many successful detections during the defense process occurring in regions closer to the target vehicle, where slight changes in angle have a greater impact on perception.

Table 9: Defense performance against camera-only attacks.

YOLOv3		Fast R-CNN	
DR(%)	RD(m)	DR(%)	RD(m)
86.0	12.8	79.3	12.7

Table 10: Defense performance against radar-only attacks.

Strong reflection		Medium reflection		Weaker reflection	
DR(%)	RD(m)	DR(%)	RD(m)	DR(%)	RD(m)
99	18.6	100	16.3	98	17.2

7.4 Defense Performance on Camera and Radar Perception

In this section, we evaluate our defense against camera-only and radar-only systems. While camera-based attacks have been widely studied, achieving a perfect physical-world attack remains challenging. Factors such as RGB distortion and the trade-off between attack effectiveness and stealth (e.g., patch size) limit their consistency across varying distances. To evaluate whether our defense can counter camera-only attacks, we deploy two adversarial boards in the CARLA simulator and conduct attacks in 10 scenarios. The initial distance is set to be 30 m. Other settings follow Table 4. The defense performance is reported in Table 9 without considering uncertainty-aware filtering at 15 km/h. The results show that even at short distances, our defense remains effective.

For the evaluation of radar-based object detection, we consider an AV equipped with radar and use the Constant False Alarm Rate (CFAR) algorithm for object detection. We consider the attack model described in [58]. In this scenario, the attacker places a smooth metal surface between the radar and the target vehicle to deflect transmitted mmWave signals away from the radar receiver. As a result, the energy of the echo signals from the target vehicle is reduced, leading to missed detections. In our experiment, we use a Honda sedan as the target vehicle. For the adversarial object, we use a 0.8 m × 0.45 m metal plate, which is placed at the rear of the target vehicle. We select three different metal plates with varying reflectivity levels: strong, medium, and weak reflection. The position and angle of each metal plate are adjusted to maximize the impact on object detection performance. The victim AV is positioned between 30 m and 9 m across seven different directions on the road segment to evaluate object detection performance under varying conditions. For each metal plate, we perform 100 defense trials and record the average DR and RD without considering uncertainty-aware filtering. The results in Table 10 show that the radar-based attacks can be significantly mitigated with our motion planning model. Besides, we can observe that even with a short initial distance, the remaining distance is much larger compared with LiDAR-only or camera-only attacks. For example, when the speed is 15 km/h, the defense only takes an average distance of 12.6 m to escape from the attack.



Figure 10: Real-world perception testbed.

7.5 Defending against Vehicle Creating Attacks

Next, we discuss how to extend the defense to address the vehicle creating attack, which aims to make the AV falsely detect a non-existent vehicle by placing adversarial objects in the environment. To the best of our knowledge, only one adversarial object-based vehicle creating attack exists [26]. In this attack, the attacker places two billboards on the roadside and injects malicious IEMI signals into the LiDAR to perturb the point cloud. As with hiding attacks, deciding when to activate the defense is crucial. For hiding attacks, the defense is triggered by sparse but sporadic detections. For vehicle creating attacks, simply inverting this, triggering defense on sporadic missed detections, can lead to unnecessary activations, as detection models may miss a few frames even in clean scenarios. However, our investigation shows that although the attack can achieve a high success rate, the “ghost car” typically has lower confidence than in clean scenarios. For example, in the clean case in Figure 2, most frames show confidence above 0.9, while the “ghost car” is consistently lower. Therefore, we adopt a dual-condition activation strategy: (1) sporadic missed detections, and (2) average confidence significantly lower than clean cases. We use 3 (the number of miss-detected frames) and 0.3 (the reduced confidence) as thresholds in our case study. The detection reward in Section 6.2 is modified to $r_t^d = -\mu_1(\gamma p_t) + \mu_2(1 - c_t) - \mu_3\phi(\Delta c_t)$ to adapt to the creating attack. In practice, the defender can switch between defense modes (for hiding or creating attacks) based on the activation condition. Following the attack setting in [26], we evaluate 30 successful attack trials with an initial distance of 30 m. Results show that our defense detects the attack in all cases, with an average RD of 15.4 m. This case study reveals that both hiding and creating attacks are sensitive to viewing angle and distance deviations.

8 Real-World Evaluation

In this section, we evaluate our defense mechanism using a real-world perception testbed for autonomous driving, as shown in Figure 10. The testbed features a Velodyne VLP-32C LiDAR sensor and a ZED 2i camera, both commonly used in AVs. Mounted on top of the vehicle, the LiDAR is positioned at 1.8 m above the ground, while the camera is at 1.7 m. The LiDAR has 32 channels with a 40° vertical field of view, and the ZED 2i camera captures high-resolution images with stable exposure. We consider AdvLoc for LiDAR-only detection and AdvFusion for LiDAR-camera fusion, using PointPillar and YOLOv3 as the object detection models.

The evaluation follows a hybrid approach, combining real-world sensory data with agent control in the CARLA simulator. We first implement AdvLoc and AdvFusion attacks using cardboard structures and printed color patches as adversarial objects, with a sedan

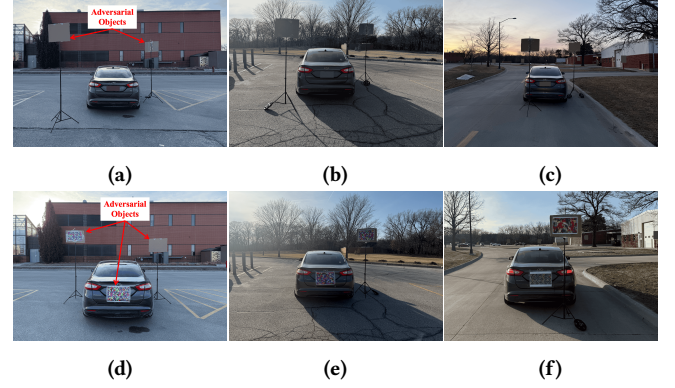


Figure 11: Real-world attack scenarios in daytime. The first row shows the AdvLoc attack, while the second row shows the AdvFusion attack. (a), (b), and (c) correspond to scenes 1, 2, and 3 for AdvLoc, while (d), (e), and (f) represent the same scenes for AdvFusion.

Table 11: Defense performance on real-world attacks.

Attack method	Scene 1		Scene 2		Scene 3	
	DR(%)	RD(m)	DR(%)	RD(m)	DR(%)	RD(m)
AdvLoc	88.3	13.5	85.0	13.1	91.7	15.0
AdvFusion	93.3	15.8	81.7	12.4	86.7	14.4

as the target vehicle. Sensory data is collected along a 40 m × 3.7 m road segment to derive detection confidence at various locations. These confidence scores are then used as inputs for the motion planning model in CARLA. During attack detection, we replace the simulator-generated confidence scores with real-world values from corresponding locations. Specifically, at time step t , the agent’s confidence score (c_t) and confidence change (Δc_t) are substituted with values from physical-world data. This hybrid approach allows for realistic defense evaluation while ensuring safety during testing.

Figure 11 illustrates the attack scenarios used in our experiment across three scenes. The first row shows the AdvLoc attack, while the second depicts AdvFusion in the same scenes. For each attack, we conduct 60 defense trials, randomly selecting the victim AV’s starting location within 35–40 m from the target vehicle. The motion planning models used for both attacks are the same as those for Table 4. Table 11 reports the average DR and RD after applying uncertainty-aware filtering. The results show that our defense mechanism consistently achieves high DR and RD across different real-world scenarios, despite variations between scenes.

To further evaluate our defense mechanism, we analyze the impact of lighting conditions and a passing vehicle. As shown in Figure 12a and Figure 12b, we launch AdvLoc and AdvFusion attacks in Scene 1 at night and test AdvLoc with a passing vehicle (Figure 12c). For each scenario, we conduct 60 defense trials, randomly selecting the testbed’s starting location within 35–40 m of the target vehicle. Table 12 reports the average results at AV speeds of 10 km/h and 15 km/h after defense activation. The results show minimal impact

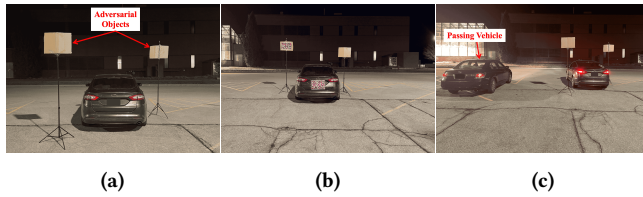


Figure 12: Real-world attack scenarios at night. (a) and (b) depict AdvLoc and AdvFusion attacks, respectively. (c) shows the AdvLoc attack with a passing vehicle.

Table 12: Impact of lighting conditions and passing vehicle.

Speed	Night-LiDAR		Night-Fusion		Passing vehicle	
	DR(%)	RD(m)	DR(%)	RD(m)	DR(%)	RD(m)
10 km/h	91.7	13.1	80.0	13.5	90.0	16.0
15 km/h	86.7	12.8	76.7	11.1	81.7	12.1

of lighting conditions on LiDAR-only detection. However, at night, LiDAR-camera fusion sees a slight performance drop, likely due to reduced lighting affecting the camera. Additionally, the passing vehicle does not impact defense performance.

9 Conclusion

In this paper, we propose a novel defense mechanism that leverages intelligent motion planning to counter adversarial object-based attacks in autonomous driving. Our approach integrates real-time attack detection with adaptive trajectory planning, guiding the AV toward regions where attacks are less effective. It employs a DRL-based motion planning model to dynamically adjust the AV's trajectory and incorporates an uncertainty-aware filtering scheme to enhance detection robustness. Extensive evaluations in both simulated and real-world environments demonstrate its effectiveness across different sensor modalities and multi-sensor fusion while ensuring safe and smooth driving.

Acknowledgments

This work was supported in part by the US National Science Foundation under grant CNS-2154059, CNS-2120369, and CNS-2413876.

References

- [1] [n. d.]. Autoware Foundation. <https://www.autoware.org/>
- [2] [n. d.]. Baidu Apollo. <https://developer.apollo.auto/>
- [3] [n. d.]. Braking distance. https://en.wikipedia.org/wiki/Braking_distance
- [4] Il Bae, Jaeyoung Moon, Junekyo Chung, Ho Suk, Taewoo Kim, Hyungbin Park, Jaekwang Cha, Jinhyuk Kim, Dohyun Kim, and Shiho Kim. 2020. Self-driving like a human driver instead of a robot: Personalized comfortable driving experience for autonomous vehicles. *arXiv preprint arXiv:2001.03908* (2020).
- [5] Stefania Bartoletti, Henk Wymeersch, Tomasz Mach, Oliver Brunnegård, Domenico Giustiniano, Peter Hammarberg, Musa Furkan Keskin, Jesus O Lacruz, Sara Modarres Razavi, Joakim Rönblom, et al. 2021. Positioning and sensing for vehicular safety applications in 5G and beyond. *IEEE Communications Magazine* 59, 11 (2021), 15–21.
- [6] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. 2021. 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)*. IEEE, 176–194.
- [7] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Ram-pazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. 2019. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2267–2281.
- [8] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. 2019. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I* 18. Springer, 52–68.
- [9] Zhiyuan Cheng, Hongjun Choi, James Liang, Shiwei Feng, Guan hong Tao, Dongfang Liu, Michael Zuzak, and Xiangyu Zhang. 2023. Fusion is Not Enough: Single Modal Attacks on Fusion Models for 3D Object Detection. *arXiv preprint arXiv:2304.14614* (2023).
- [10] Pranav Singh Chib and Pravendra Singh. 2023. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Transactions on Intelligent Vehicles* (2023).
- [11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*. PMLR, 1–16.
- [12] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1625–1634.
- [13] Ross Girshick. 2015. Fast r-cnn. *arXiv preprint arXiv:1504.08083* (2015).
- [14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
- [15] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. 2020. A survey of deep learning techniques for autonomous driving. *Journal of field robotics* 37, 3 (2020), 362–386.
- [16] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. Badnets: Evaluating backdoor attacks on deep neural networks. *IEEE Access* 7 (2019), 47230–47244.
- [17] Zhuangzhuang Gu, Hem Regmi, and Sanjib Sur. 2023. Poster: mmBox: mmWave Bounding Box for Vehicle and Pedestrian Detection Under Outdoor Environment. In *2023 IEEE 31st International Conference on Network Protocols (ICNP)*. IEEE, 1–2.
- [18] Junfeng Guan, Sohrab Madani, Suraj Jog, and Haitham Hassanieh. 2019. High resolution millimeter wave imaging for self-driving cars. *arXiv preprint arXiv:1912.09579* (2019).
- [19] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. Pmlr, 1861–1870.
- [20] Jun Han, Albert Jin Chung, Manal Kumar Sinha, Madhumitha Harishankar, Shijia Pan, Hae Young Noh, Pei Zhang, and Patrick Tague. 2018. Do you feel what I hear? Enabling autonomous IoT device pairing using different sensor types. In *2018 IEEE symposium on Security and Privacy (SP)*. IEEE, 836–852.
- [21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [22] Zizhi Jin, Xiaoyu Ji, Yushi Cheng, Bo Yang, Chen Yan, and Wenyuan Xu. 2023. Plalidar: Physical laser attacks against lidar-based 3d object detection in autonomous vehicle. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1822–1839.
- [23] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7122–7131.
- [24] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12697–12705.
- [25] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2022. Backdoor learning: A survey. *IEEE transactions on neural networks and learning systems* 35, 1 (2022), 5–22.
- [26] Ziwei Liu, Feng Lin, Teshi Meng, Benaouda Chouaib Baha-eddine, Li Lu, Qiang Xue, and Kui Ren. 2024. EMTrig: Physical Adversarial Examples Triggered by Electromagnetic Injection towards LiDAR Perception. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 351–364.
- [27] Yongsan Ma, Gang Zhou, and Shuangquan Wang. 2019. Wi-Fi sensing with channel state information: A survey. *ACM Computing Surveys (CSUR)* 52, 3 (2019), 1–36.
- [28] Aleksander Madry. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).
- [29] Delton Martin and Dale Litwhiler. 2008. An investigation of acceleration and jerk profiles of public transportation vehicles. In *2008 Annual Conference & Exposition*. 13–194.
- [30] John Nolan, Kun Qian, and Xinyu Zhang. 2021. RoS: passive smart surface for roadside-to-vehicle communication. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*. 165–178.
- [31] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Niwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. 2023. Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 530–543.

- [32] Akarsh Prabhakara, Tao Jin, Arnav Das, Gantavya Bhatt, Lilly Kumari, Elahe Soltanaghai, Jeff Bilmes, Swarun Kumar, and Anthony Rowe. 2023. High resolution point clouds from mmwave radar. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4135–4142.
- [33] Akarsh Prabhakara, Diana Zhang, Chao Li, Sirajum Munir, Aswin C Sankaranarayanan, Anthony Rowe, and Swarun Kumar. 2022. Exploring mmWave Radar and Camera Fusion for High-Resolution and Long-Range Depth Imaging. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3995–4002.
- [34] Kun Qian, Shilin Zhu, Xinyu Zhang, and Li Erran Li. 2021. Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 444–453.
- [35] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.
- [36] Sergio Alberto Salguero-Luna, Kelsey Alejandra Ramirez-Gutiérrez, and Alfonso Martinez-Cruz. 2024. A State-of-the-Art Review on Attacks and Defense Mechanisms for LiDAR on Autonomous Vehicles. *IEEE Transactions on Intelligent Transportation Systems* (2024).
- [37] Louis I. Scharf and Cédric Demeure. 1991. Statistical signal processing: detection, estimation, and time series analysis. (No Title) (1991).
- [38] Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research* 9, 3 (2008).
- [39] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 2019. Pointtrcn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 770–779.
- [40] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. 2021. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing* 107 (2021), 104117.
- [41] Jiachen Sun, Yulong Cao, Qi Alfred Chen, and Z Morley Mao. 2020. Towards robust {LiDAR-based} perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *29th USENIX Security Symposium (USENIX Security 20)*. 877–894.
- [42] Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Yunfeng Ai, Dongsheng Yang, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, et al. 2023. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles* 8, 6 (2023), 3692–3711.
- [43] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2017. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204* (2017).
- [44] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. 2020. Physically realizable adversarial examples for lidar object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13716–13725.
- [45] Ningfei Wang, Yunpeng Luo, Takami Sato, Kaidi Xu, and Qi Alfred Chen. 2023. Does physical adversarial example really matter to autonomous driving? towards system-level effect of adversarial object evasion attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4412–4423.
- [46] Xuan Wang, Kaiqiang Li, and Abdellah Chehri. 2023. Multi-sensor fusion technology for 3D object detection in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems* (2023).
- [47] Hui Wei, Hao Tang, Xuemei Jia, Zhixiang Wang, Hanxun Yu, Zhubo Li, Shin'ichi Satoh, Luc Van Gool, and Zheng Wang. 2024. Physical adversarial attack meets computer vision: A decade survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [48] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. 2017. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*. 1369–1378.
- [49] Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. 2022. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. *Advances in Neural Information Processing Systems* 35 (2022), 25667–25682.
- [50] Bin Yang, Wenjie Luo, and Raquel Urtasun. 2018. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 7652–7660.
- [51] Jiadi Yu, Hongzi Zhu, Haofu Han, Yingying Jennifer Chen, Jie Yang, Yanmin Zhu, Zhongyang Chen, Guangtao Xue, and Minglu Li. 2015. Senspeed: Sensing driving conditions to estimate vehicle speed in urban environments. *IEEE Transactions on Mobile Computing* 15, 1 (2015), 202–216.
- [52] Jinghui Zhang, Jianfeng Chi, Zheng Li, Kunlin Cai, Yang Zhang, and Yuan Tian. 2024. Badmerging: Backdoor attacks against model merging. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*. 4450–4464.
- [53] Yan Zhang, Zihao Liu, Chongliu Jia, Yi Zhu, and Chenglin Miao. 2024. An Online Defense against Object-based LiDAR Attacks in Autonomous Driving. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 380–393.
- [54] Yan Zhang, Zihao Liu, Yi Zhu, and Chenglin Miao. 2025. Towards Real-Time Defense against Object-Based LiDAR Attacks in Autonomous Driving. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*. 3825–3839.
- [55] Yan Zhang, Yi Zhu, Zihao Liu, Chenglin Miao, Foad Hajiaghajani, Lu Su, and Chunming Qiao. 2022. Towards backdoor attacks against lidar object detection in autonomous driving. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 533–547.
- [56] Shenchen Zhu, Yue Zhao, Kai Chen, Bo Wang, Hualong Ma, et al. 2024. AE-Morpher: Improve Physical Robustness of Adversarial Objects against LiDAR-based Detectors via Object Reconstruction. In *33rd USENIX Security Symposium (USENIX Security 24)*. 7339–7356.
- [57] Yi Zhu, Chenglin Miao, Hongfei Xue, Zhengxiong Li, Yunnan Yu, Wenyao Xu, Lu Su, and Chunming Qiao. 2023. TileMask: A Passive-Reflection-based Attack against mmWave Radar Object Detection in Autonomous Driving. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 1317–1331.
- [58] Yi Zhu, Chenglin Miao, Hongfei Xue, Yunnan Yu, Lu Su, and Chunming Qiao. 2024. Malicious Attacks against Multi-Sensor Fusion in Autonomous Driving. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 436–451.
- [59] Yi Zhu, Chenglin Miao, Tianhang Zheng, Foad Hajiaghajani, Lu Su, and Chunming Qiao. 2021. 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*. 1945–1960.