Fooling Local Optimizers: Adversarial Initialization Attacks on LiDAR Localization

Madhurima Ghosh
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
madhurima.ghosh@cispa.de

Pranav Shetty
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
pranav.shetty@cispa.de

Hyungsub Kim
Indiana University
Bloomington, United States
hk145@iu.edu

Mridula Singh
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
singh@cispa.de

Abstract

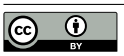
LiDAR (light detection and ranging) localization is a key component of autonomous driving systems, enabling accurate vehicle pose estimation through optimization-based scan registration methods (e.g., the normal distributions transform algorithm). Existing adversarial attacks mainly target the observation space by manipulating LiDAR point clouds, overlooking another important dependency: the position prior used to initialize the optimizer. In this paper, we identify the initialization space as a previously unexplored attack surface and show that GNSS spoofing can drive LiDAR localization to incorrect local optima. Experiments on the Autoware-CARLA platform reveal a convergence basin of approximately **4 m**, beyond which initialization perturbations lead to localization failures. Our analysis of observation-space attacks shows that localization errors induced by LiDAR spoofing exhibit a stable magnitude and a dominant drift direction over time. Leveraging these insights, we propose a cross-space adversarial attack that jointly exploits the initialization and observation spaces by coordinating GNSS spoofing with LiDAR manipulation, demonstrating that coordinated perturbations across sensing modalities can propagate through the complete localization stack. The proposed attack increases the mean localization error by approximately **7 m** for A-LOAM and **15 m** for HDL-Localizer compared to observation-space attacks alone. These findings expose vulnerabilities arising from interactions between initialization and observation spaces and provide a basis for evaluating the security of autonomous vehicle localization systems.

CCS Concepts

• **Computer systems organization** → **Embedded and cyber-physical systems**; **Sensors and actuators**; **Robotics**; **Sensor networks**.

Keywords

LiDAR localization, adversarial attacks, initialization space, GNSS spoofing, LiDAR manipulation, local optimization



AISec '26, The Hague, Netherlands

ACM ISBN 979-8-4007-3031-3/2026/11
https://doi.org/10.1145/3847352.3848085

ACM Reference Format:
Madhurima Ghosh, Pranav Shetty, Hyungsub Kim, and Mridula Singh. 2026. Fooling Local Optimizers: Adversarial Initialization Attacks on LiDAR Localization. In *19th Workshop on Artificial Intelligence and Security (AISec '26), November 15–19, 2026, The Hague, Netherlands.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3847352.3848085

1 Introduction

Precise localization is fundamental to autonomous vehicles (AVs), as accurate vehicle position estimates are essential for safe navigation and decision-making in complex environments. To achieve robust localization, AVs commonly employ multi-sensor fusion (MSF) [2, 17], which combines complementary measurements from global navigation satellite systems (GNSS), inertial measurement units (IMUs), light detection and ranging (LiDAR), and other sensors.

Among these modalities, LiDAR-based localization has gained widespread adoption in AVs because LiDAR provides dense and geometrically informative measurements that enable accurate position estimation relative to a reference map. However, unlike GNSS and IMUs, which provide direct positional or motion measurements, LiDAR localization (as shown in Figure 1) is a machine learning-based localization model that estimates the vehicle position through iterative optimization. Given a LiDAR point cloud and an initial position estimate (position prior), the localization model iteratively registers the observation against a reference map to refine the vehicle position. Representative approaches, including iterative closest point (ICP) [18], normal distributions transform (NDT) [19], LiDAR odometry and mapping (LOAM) [28], and their variants [5, 25], perform this inference through nonlinear optimization.

Failures in LiDAR localization can compromise the reliability of the overall localization MSF, particularly when sensor measurements are corrupted or manipulated by an adversary. Attacks can target individual sensing modalities or exploit the fusion process to induce erroneous vehicle position estimates. Moreover, coordinated attacks across multiple sensors can simultaneously manipulate complementary measurements, potentially making malicious estimates more consistent with each other and harder for the MSF to detect. Such localization attacks can propagate to downstream autonomous driving software modules, leading to incorrect trajectory planning, degraded obstacle avoidance, mapping errors, and potentially unsafe vehicle behavior. Therefore, understanding the impact of attacks on LiDAR localization, as well as coordinated

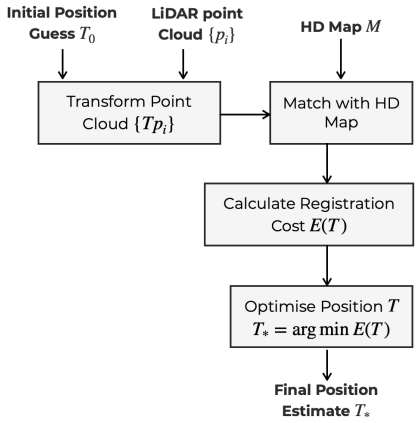


Figure 1: Overview of LiDAR localization through position optimization. Previous works exploit the LiDAR point cloud (observation space), whereas we target the initial position guess through GNSS spoofing.

attacks on the broader localization MSF, is critical for developing secure and reliable autonomous vehicles (AVs).

Recent studies [12, 13, 20] have investigated the vulnerability of LiDAR localization to adversarial attacks by manipulating physical LiDAR observations. These approaches primarily operate in the observation space, for example, by inserting, removing, or modifying objects in the LiDAR point cloud to mislead scan-matching algorithms [18, 19]. While these attacks demonstrate vulnerabilities in LiDAR localization, two important aspects remain largely unexplored.

First, existing attacks primarily target LiDAR localization in isolation and do not examine their impact on the broader localization MSF. In commercial AV systems such as Apollo [2], Waymo [7], and Autoware [21], LiDAR localization is coupled with GNSS and IMUs for position estimates. Thus, an attack that corrupts LiDAR localization may not necessarily corrupt the final MSF estimate. A successful attack must account for these sensor fusion pipelines and potentially coordinate with other sensor measurements to overcome the redundancy of MSF.

Second, existing works mainly focus on direct point-cloud manipulation. This can be challenging in modern LiDAR sensors such as Hesai, Ouster, Robosense, where timing randomization mechanisms complicate precisely timed point injection [15], while physical object removal may require specialized hardware [13]. This motivates investigating attack surfaces beyond direct point-cloud manipulation, particularly those arising from the dependencies between LiDAR localization and the localization conducted by the MSF.

To address these gaps, we investigate the security of LiDAR localization and analyze existing observation-space attacks. We further identify a previously unexplored attack surface in LiDAR localization: the initialization space. Instead of directly modifying LiDAR point clouds, we demonstrate that an adversary can manipulate the position prior (initial guess) and lead the optimizer to an incorrect local minimum to produce a localization failure. Finally, we investigate coordinated attacks that jointly manipulate LiDAR and GNSS measurements. This enables us to evaluate LiDAR localization not only in isolation, but also within the MSF-based localization pipeline used by AVs.

In summary, we make the following contributions:

- **Analysis of observation-space attacks.** We systematically study existing attacks that manipulate the LiDAR point cloud and analyze their impact on optimization-based LiDAR localization.
- **Identification of initialization space as a new attack surface.** We identify the position prior used to initialize LiDAR localization as a previously unexplored attack surface and demonstrate how an adversary can exploit it in AV localization systems through GNSS spoofing.
- **Cross-space coordinated attacks.** We develop coordinated attacks that simultaneously manipulate LiDAR observations and GNSS-based position priors to maximize localization error. By jointly exploiting the observation and initialization spaces, these attacks demonstrate how an adversary can overcome the redundancy of localization MSF and induce substantially larger localization errors.

2 Background and Related Works

2.1 LiDAR Localization

LiDAR-based localization is a fundamental machine learning (ML) problem in autonomous driving and robotic systems, where models learn to estimate the sensor's position by aligning LiDAR observations with either previous observations or a pre-built representation of the environment. Existing LiDAR localization approaches can be broadly categorized into two classes: scan-to-scan and scan-to-map localization.

Scan-to-scan localization [25, 28] estimates the relative motion between consecutive LiDAR scans by registering the current observation against the previous scan. These approaches are commonly used in simultaneous localization and mapping (SLAM) [5] systems for incremental position estimation and map construction. However, since each position estimate depends on previous estimates, scan-to-scan approaches are susceptible to accumulated drift over long trajectories.

In contrast, scan-to-map localization [11] estimates the vehicle position by aligning the current LiDAR scan against a pre-built map. This approach is commonly adopted in commercial autonomous driving systems due to its ability to provide globally consistent localization. By leveraging environmental maps, scan-to-map methods can reduce long-term drift compared with scan-to-scan approaches.

Regardless of the localization paradigm, most LiDAR localization systems formulate localization as a point cloud registration problem. Representative registration algorithms include iterative closest point (ICP) [18], normal distributions transform (NDT) [19], and LiDAR odometry and mapping (LOAM) [28]. These algorithms typically solve nonlinear optimization problems using local optimization procedures. Consequently, their convergence depends heavily on the quality of the initial position estimate. If the initial guess is sufficiently close to the true position, the optimizer can converge to the correct solution; however, large initialization errors can cause convergence to incorrect local optima.

2.2 Multi-Sensor Fusion for Localization

Commercial autonomous driving systems rely on multi-sensor fusion (MSF) modules [2, 17] to achieve robust localization and navigation. Sensor redundancy is introduced by combining complementary sensing modalities, where GNSS provides global positioning information, IMU provides motion constraints, and LiDAR localization provides geometric corrections from environmental observations.

Typically, these measurements are fused asynchronously using Kalman filters or their variants, such as Extended Kalman Filters (EKF) [26, 30] and Error-State Kalman Filters (ESKF) [1, 9]. The fusion framework maintains an estimate of the vehicle state and continuously updates it based on incoming sensor measurements. By exploiting redundancy among sensors, these systems can compensate for individual sensor failures and improve localization robustness.

Prior works have investigated the security of MSF-based localization systems. However, existing studies have primarily focused on extracting or modeling the behavior of the underlying Kalman filter [29], or have considered attacks targeting individual sensors, particularly GNSS spoofing [16, 27]. These approaches often overlook the interaction between sensor fusion and downstream LiDAR localization optimization.

Different autonomous driving platforms employ different localization architectures. For example, systems such as Autoware [21] primarily rely on LiDAR localization and use GNSS mainly to provide the initial localization estimate. In contrast, commercial autonomous driving platforms such as Baidu Apollo [2] and Waymo [7] integrate GNSS, IMU, and LiDAR localization measurements directly into the fusion framework. These architectural differences introduce different attack surfaces and motivate the need to analyze localization security from a system-level perspective.

2.3 Attacks on LiDAR Localization

Recent works have demonstrated the vulnerability of LiDAR localization to physical adversarial attacks, primarily by manipulating the LiDAR observation space. Nagata et al. [13] introduced SLAMSpoof, a framework for physically realizable LiDAR spoofing attacks based on Chosen Point Injection (CPI) [3] and High Frequency Removal (HFR) [15]. CPI injects fabricated LiDAR points into the point cloud, whereas HFR removes legitimate LiDAR pulse reflections by exploiting the physical characteristics of fired pulses from LiDAR sensors. HFR is particularly relevant to modern LiDARs, such as Livox, Ouster, and RoboSense, which employ timing randomization that mitigates injection-based attacks. Tanaka et al. [20] demonstrated that adversarial point injection can mislead the NDT localization module in Autoware, while Lao et al. [12] showed that selectively removing LiDAR points can degrade localization by disrupting point-cloud registration. In our baseline evaluation, we focus on the HFR attack (using SLAMSpoof framework) because of its practical applicability to modern LiDAR systems.

Despite these advances, existing attacks primarily treat the LiDAR point cloud as the attack surface, assuming that the localization optimizer and its initialization remain trustworthy. In this work, we extend this perspective by investigating additional attack surfaces in the localization pipeline, including the initialization provided by external localization sources and the interactions between multiple sensing modalities.

3 Observation Space Vulnerability Analysis

We conduct a systematic study of existing physical attacks on LiDAR localization and find that the attacks in the SLAMSpoof framework [13] account for the physical constraints of modern LiDARs, including Ouster, Hesai, and RoboSense sensors used in industry-grade AVs. We therefore evaluate SLAMSpoof against LiDAR localization as a baseline and analyze the magnitude and direction of the resulting localization error over time. This analysis characterizes the error induced by observation-space attacks and provides the basis for the cross-space attack strategy presented in Section 5.

3.1 Threat Model

We adopt the physical LiDAR spoofing threat model proposed by Sato et al. [15] and instantiate attacks using the SLAMSpoof [13] framework. The adversary is equipped with a spoofing device consisting of a pulse laser and an optical lens system capable of injecting forged LiDAR pulse reflections at a high firing rate into the victim vehicle's sensor. The objective of the attacker is to perturb the victim's localization estimate without compromising the onboard software stack.

Among the attack primitives provided by SLAMSpoof, we employ the HFR [15] attack as it is effective against new generation automotive LiDARs, including sensors from Ouster, Livox, Hesai, and RoboSense. In this attack, the adversary injects spoofed LiDAR pulse reflections such that legitimate LiDAR measurements within an approximately 80° azimuth sector are replaced by these fake rays. As a result, objects located within this sector are effectively removed from the LiDAR scan, producing a consistent blind region that biases the downstream localization algorithm.

However, SLAMSpoof evaluates LiDAR localization modules in isolation, without integrating the upstream multi-sensor localization framework. Consequently, the LiDAR localization algorithm does not receive position priors from external sources such as GNSS or sensor fusion. Instead, the initial position estimate for each registration is obtained from the localization estimate of the previous frame, effectively propagating the estimated trajectory through successive scan registrations. This setup enables us to isolate and analyze the effects of observation-space perturbations on LiDAR localization, independent of external initialization errors.

The victim system consists of a LiDAR-based localization module operating on sequential LiDAR scans. We evaluate two representative localization pipelines:

- A-LOAM [28], a scan-to-scan localization algorithm that estimates motion by registering consecutive LiDAR scans.
- HDL Localizer [11], a scan-to-map localization algorithm that estimates position by registering the current scan against a pre-built map using Normal Distributions Transform (NDT).

All experiments are performed using sequences from the KITTI dataset [8], with spoofed LiDAR observations generated through the SLAMSpoof framework. Localization error is evaluated by comparing the estimated trajectory against the ground-truth positions provided by the dataset.

3.2 Analysis of Induced Localization Error

We first reproduce the results of prior work [13] and establish its performance as a baseline for our evaluation. We then conduct multiple tests to analyze the temporal evolution and directionality of the induced localization errors, as discussed below. These characteristics provide insights into the behavior of optimization-based localization under adversarial observations and motivate our subsequent initialization attacks.

3.2.1 Error Magnitude Across Localization Models. We first compare the localization errors produced by HFR attacks on representative scan-to-scan and scan-to-map localization algorithms.

Model	Position Error (m)			Orientation Error (°)		
	Mean	Max	SD	Mean	Max	SD
A-LOAM [28]	5.72	19.15	4.64	1.70	13.82	1.65
HDL Localizer [11]	0.01	3.06	0.05	1.15	4.20	1.33

Table 1: Localization Error Evaluation for Baseline Models

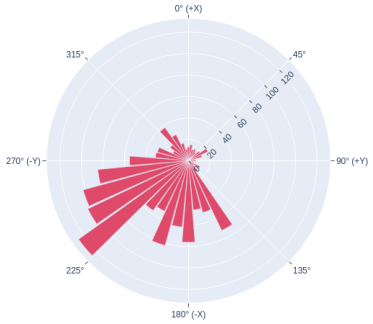
Table 1 summarizes the localization errors obtained for A-LOAM and HDL-Localizer. Consistent with the observations reported in SLAMSpoof, scan-to-scan localization exhibits significantly larger localization errors than scan-to-map localization. Since scan-to-scan methods estimate vehicle motion by registering consecutive LiDAR scans, errors introduced by spoofed observations directly affect the relative motion estimate and accumulate over time. In contrast, scan-to-map localization continuously aligns the current scan with a globally consistent map, providing additional geometric constraints that partially mitigate the effect of spoofed observations.

These results validate previous findings while confirming that map-based localization is inherently more resilient to observation-space perturbations than scan-to-scan approaches.

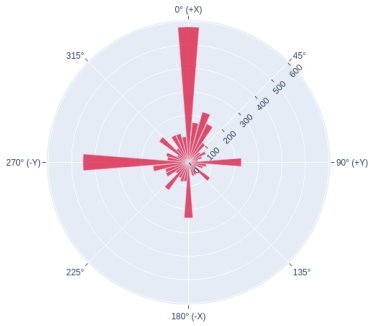
3.2.2 Directionality of Localization Drift. The localization error induced by physical LiDAR spoofing is not deterministic across individual frames. Unlike GNSS, LiDAR point clouds do not directly provide a positional measurement; the estimated pose is instead obtained as the solution of an optimization problem. Consequently, small variations in the observed point cloud can lead to variations in the resulting localization error.

However, closer examination shows that the error direction remains relatively consistent across consecutive observations. Figure 2 shows the distribution of localization error directions for A-LOAM and HDL-Localizer. Although the error direction may appear random, most of the consecutive error vectors are concentrated within a narrow range of directions. Individual localization errors therefore vary in magnitude and direction, but their directions remain correlated over time. This correlation causes the errors to accumulate along consistent directions, resulting in a systematic drift.

To quantify this behavior, we compute the standard deviation of the localization error magnitude and direction over five consecutive frames, with the results summarized in Table 2. The low directional variability confirms that consecutive errors are correlated rather than independent. This persistent directional bias causes localization errors to accumulate over time.



(a) A-LOAM



(b) HDL-Localizer

Figure 2: Distribution of localization error directions for A-LOAM and HDL-Localizer. Although individual error directions may vary, the error vectors exhibit clear directional concentration, with a large fraction of errors accumulating along specific directions rather than being uniformly distributed.

LiDAR Localization Model	Window Type	Direction of Error (°)		Magnitude of Error (m)	
		Mean	SD	Mean	SD
A-LOAM [28]	Non-overlapping	7.60	8.74	0.321	0.211
	Overlapping	7.82	8.70	0.319	0.213
HDL Localizer [11]	Non-overlapping	12.19	11.53	0.103	0.222
	Overlapping	12.35	11.70	0.104	0.224

Table 2: Position Error comparison of LiDAR localization models showing direction and magnitude of error across window types (Window size = 5).

3.2.3 Repeatability of Localization Drift. To evaluate whether the observed error behavior depends on the placement of the spoofing device, we conduct repeatability experiments by varying its position. For each localization algorithm, we perform the HFR attack with six different spoofer placements that exhibit a high Scan Matching Vulnerability Score (SMVS) [13]. This experimental setup allows us to assess the consistency of the observed localization errors for various spoofer placements.

Across the resulting 12 experimental runs, 11 runs exhibit consistent localization behavior. Specifically, provided that the spoofing device is positioned at a location with high SMVS, the induced

localization trajectories follow a similar drift pattern, with the localization error varying by less than 1 m across repeated experiments. Figure 3 illustrates a representative trajectory for A-LOAM, where the localization error accumulates progressively over time as successive spoofed observations bias the optimization process.

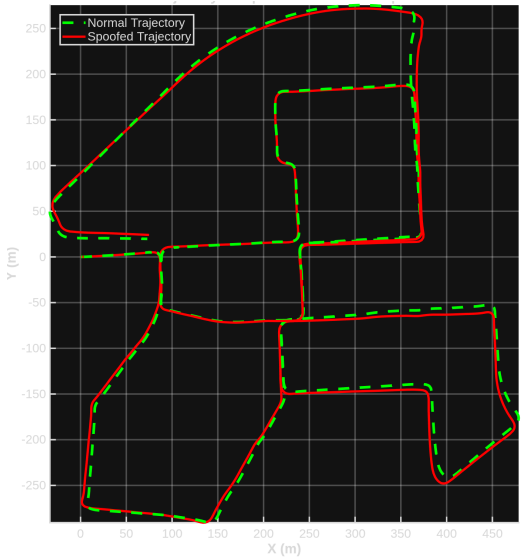


Figure 3: Accumulated Trajectory Drift due to LiDAR spoofing on A-LOAM.

These results indicate that the effectiveness of HFR attacks is largely independent of the exact spoofing location, provided the calculated SMVS of the chosen location is high. More importantly, the consistent drift observed across repeated experiments suggests that the localization error is a systematic consequence of the optimization process rather than an artifact of a particular attack configuration.

3.2.4 Convergence Behavior of Local Optimization. While the localization trajectories remain consistent in 11 of the 12 experiments, we observe a qualitatively different behavior in one experiment involving HDL-Localizer. This failure mode was not reported in the original SLAMSpoof study and was identified in our evaluation.

Instead of exhibiting a gradual accumulation of localization error, the estimated vehicle position eventually loses localization consistency and remains confined to a small region despite continued vehicle motion. Figure 4 illustrates this behavior, where the estimated position effectively hovers around a fixed location while the ground-truth vehicle continues along its trajectory.

We attribute this phenomenon to the convergence characteristics of the underlying LiDAR Localization. Under prolonged spoofing, the optimizer converges to an incorrect local minimum from which it is unable to recover. Since the SLAMSpoof framework evaluates LiDAR localization modules in isolation, the initial position supplied to each optimization iteration is simply the localization estimate from the previous frame. Once the optimizer becomes trapped in an incorrect basin of attraction, the erroneous position estimate is

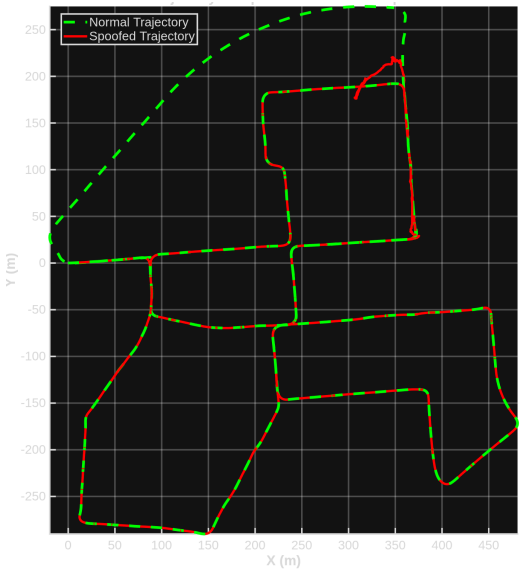


Figure 4: Comparison of Trajectory of HDL-Localizer after LiDAR Spoofing.

repeatedly propagated as the initialization for subsequent scan registrations. Consequently, each successive optimization converges to the same incorrect solution, preventing the localization module from recovering and causing the estimated vehicle position to remain effectively stationary.

Although this behavior was observed in only one experimental run, it provides important evidence that optimization-based LiDAR localization is fundamentally dependent on initialization quality. Observation-space perturbations can not only induce gradual localization drift but can also drive the optimizer into an incorrect local minimum, from which recovery is difficult without an external correction. This observation further motivates our investigation of initialization-space attacks, where the position prior itself is manipulated to intentionally steer the optimizer toward incorrect convergence.

3.3 Key Findings

Our analysis reveals two important characteristics of optimization-based LiDAR localization under observation-space attacks. First, physical LiDAR spoofing introduces a persistent directional bias, causing localization errors to accumulate gradually rather than manifest as isolated deviations. Second, prolonged observation-space perturbations can drive local optimization algorithms into incorrect basins of attraction, from which they are unable to recover because each optimization iteration is initialized using the previous localization estimate.

These findings suggest that the initialization supplied to the optimizer is itself a critical attack surface. If observation-space perturbations alone can eventually force the optimizer into an incorrect local minimum, then directly manipulating the optimizer's initial position estimate should allow an adversary to induce similar failures more efficiently.

4 Initialization Attack

We now investigate a complementary attack surface: the initialization of the optimization process. Since iterative registration algorithms converge to local optima based on an initial position estimate, an adversary capable of manipulating this initialization can influence localization without modifying the LiDAR observations.

To evaluate this vulnerability, we consider GNSS spoofing as a practical mechanism for manipulating the initial position estimate supplied to the LiDAR localization module.

LiDAR localization (as shown in Figure 1) can be formulated as the following optimization problem,

$$\hat{\mathbf{T}}_t = \arg\min_{\mathbf{T}} \mathcal{L}\left(\mathcal{M}, \mathcal{P}_t; \mathbf{T}, \mathbf{T}_t^0\right), \tag{1}$$

where $\mathcal{M}$ denotes the reference map (or previous scan for scan-to-scan methods), $\mathcal{P}_t$ denotes the current LiDAR observation, $\mathbf{T}_t^0$ represents the initial position estimate provided to the optimizer, and $\hat{\mathbf{T}}_t$ is the optimized vehicle position.

Existing LiDAR localization attacks primarily manipulate the observation space by perturbing $\mathcal{P}_t$. In contrast, we target the initialization space by introducing an adversarial perturbation to the position prior,

$$\tilde{\mathbf{T}}_t^0 = \mathbf{T}_t^0 + \delta_{\text{GNSS}}, \tag{2}$$

where δ_{GNSS} denotes the perturbation introduced through GNSS spoofing. By providing the optimizer with an erroneous initial estimate, the adversary can steer the registration process toward an incorrect local optimum even when the LiDAR observations remain authentic.

4.1 Threat Model

We consider an adversary capable of spoofing the GNSS signals received by the victim AV. Unlike the observation-space attack presented in Section 3, the attacker does not manipulate the LiDAR point cloud or interfere with the LiDAR sensor. Instead, the objective is to perturb the position prior used to initialize LiDAR localization. Physical GNSS spoofing is a well-established attack in which an adversary broadcasts counterfeit signals that cause a victim receiver to derive an erroneous position [23]. In particular, civilian GPS signals use publicly specified pseudorandom noise (PRN) codes [10], enabling an attacker to generate signals that emulate legitimate satellite transmissions. Thus, compared with physical LiDAR spoofing, GNSS spoofing can provide a lower-cost attack surface because it does not require direct interaction with the LiDAR sensor or specialized hardware for manipulating individual LiDAR returns.

Our victim platform is Autoware [21], a widely adopted open-source autonomous driving stack. Autoware performs localization using a MSF pipeline, illustrated in Figure 5. The localization stack employs NDT for LiDAR localization. During system initialization, the first position estimate is obtained from the GNSS receiver. Subsequently, the localization module relies on an Extended Kalman Filter (EKF) prediction, which fuses multiple sensor observations to provide the initial position estimate for each successive NDT registration.

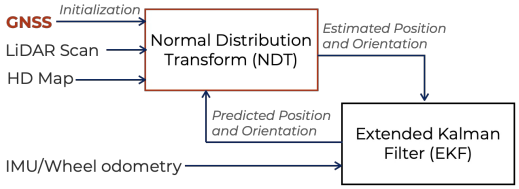


Figure 5: Threat Model for Initialization Attack. The attacker spoofs the GNSS receiver to perturb the position prior supplied to the localization MSF. During initialization, GNSS provides the initial position estimate, while subsequent position priors are generated by the EKF. The NDT LiDAR localizer refines these priors using LiDAR observations and serves as the attack target.

Consequently, the localization process can be viewed as a sequence of local optimization problems, where each optimization is initialized using the position predicted by the MSF framework. An incorrect initialization can therefore alter the optimizer's convergence behavior, even when the LiDAR observations remain unmodified.

We assume that the adversary can spoof the GNSS measurements during the system initialization phase, thereby introducing an erroneous initial position into the localization pipeline. After initialization, no further manipulation of GNSS or LiDAR measurements is performed. This threat model isolates the effect of initialization errors on the convergence behavior of the LiDAR localization algorithm.

4.2 Experimental Setup

To evaluate the sensitivity of LiDAR localization to initialization errors, we conduct experiments using Autoware in the CARLA simulation environment [4, 6]. The AV is kept stationary during each experiment to eliminate motion-induced uncertainty and isolate the effect of initialization on LiDAR localization convergence.

Before the first execution of the NDT localization module, the attacker injects a controlled offset into the GNSS-derived initial position provided to the localization system. We systematically vary three parameters:

- the magnitude of the GNSS initialization error,
- the direction of the GNSS perturbation, and
- the ground-truth position of the AV with respect to the map.

The first two parameters characterize the spoofing perturbation, while the third captures the influence of the local map geometry on the localization process. In particular, varying the AV's ground-truth position allows us to determine whether the same GNSS perturbation produces different convergence behavior at different locations in the map.

We consider localization *successful* when the final NDT estimate is within **0.6 m** of the ground-truth position and a *failure* otherwise. We adopt this conservative threshold based on prior work [16], which reports that lateral localization errors exceeding **0.895 m** can induce off-road localization and unsafe navigation behavior. We characterize the NDT *convergence basin* by performing a random search over initialization positions while varying the GNSS error

magnitude, perturbation direction, and AV ground-truth position with respect to the map. For each sampled initialization, we execute NDT and determine whether it converges to a localization estimate within the 0.6 m success threshold.

Figure 6 illustrates representative examples of the resulting localization behavior for different initialization offsets. Table 3 summarizes representative initialization configurations and their corresponding convergence outcomes.

4.3 Initialization Error Bounds

The results reveal that NDT exhibits a finite basin of convergence with respect to the initialization position, as illustrated in Figure 6. When the initialization lies within this basin, the scan-to-map registration process can recover the correct vehicle position. As the initialization is moved farther from the ground-truth position, however, the optimizer can converge to an incorrect local optimum.

For the default Autoware configuration evaluated in our experiments, the random search indicates that the convergence basin extends approximately **4 m** from the ground-truth position. This boundary represents the empirically observed spatial extent of initial positions from which NDT can recover the correct localization solution. Importantly, the convergence basin is not determined solely by the magnitude of the initialization error. Its behavior also depends on the perturbation direction and the ground-truth position of the AV with respect to the map.

Perturbation Direction	Magnitude (m)	Localization Outcome
Lateral	1.0	Successful
	2.0	Successful
	2.5	Successful
	3.0	Successful
	4.0	Failure
	5.0	Failure
	8.0	Failure
Longitudinal	2.0	Successful
	2.5	Failure
	2.5	Successful
	3.0	Successful
	4.0	Failure
	5.0	Failure
	8.0	Failure

Table 3: Localization outcomes under varying GNSS initialization perturbations. The outcome depends on the perturbation magnitude and direction, as well as the AV's ground-truth position on the map. A localization outcome is classified as *Successful* when the final localization error is less than 0.6 m from the ground-truth position, and *Failure* otherwise.

Consequently, identical GNSS error magnitudes can produce different localization outcomes at different vehicle positions or along different perturbation directions. This explains, for example, the two 2.5 m longitudinal perturbations reported in Table 3: although their perturbation magnitudes and directions are identical, their different ground-truth positions relative to the map result in different localization outcomes.

We further observe downstream consequences of the resulting localization failures for autonomous driving systems. In our experiments, incorrect convergence of the NDT optimizer caused the estimated vehicle position to deviate from the true road geometry, corresponding to an incorrect lane, a different position along the road, or even an off-road location. Such errors can propagate to downstream perception, planning, and control modules, potentially leading to unsafe driving behavior.

Figure 7 demonstrates one such consequence in Autoware. When spoofed initialization drives LiDAR localization toward an incorrect local minimum, the resulting position estimate becomes inconsistent with the vehicle's actual physical location. The navigation stack then treats this erroneous estimate as the vehicle's current position and plans the route accordingly. As a result, the vehicle followed a trajectory opposite to the intended direction and ultimately reached a different destination. Thus, initialization spoofing can extend beyond localization degradation to influence navigation decisions and produce incorrect autonomous driving outcomes.

Overall, these results indicate that NDT's convergence behavior is governed by the interaction between the initialization error and the local optimization landscape induced by the map. An initialization outside the convergence basin can drive the optimizer toward an incorrect local optimum, even when the LiDAR observations themselves remain completely authentic. This establishes the GNSS-derived initialization as a practical attack surface for LiDAR localization and demonstrates that manipulating only the initial position can have consequences that propagate from localization to autonomous navigation.

5 Cross-Space Adversarial Attack

The attacks demonstrated in the previous sections target LiDAR localization in isolation and do not directly account for the localization MSF pipeline used in autonomous driving systems. Since LiDAR localization is coupled with GNSS and IMU, compromising the LiDAR localization module alone may not be sufficient to induce a persistent error in the final localization estimate. This motivates the need for attack strategies that explicitly target the interactions within the MSF pipeline.

Building on these observations, we formalize a black-box cross-space adversarial attack that simultaneously compromises multiple sensor inputs. Specifically, the attacker jointly manipulates LiDAR observations and GNSS measurements to influence both the localization optimization and its initialization, while accounting for the consistency constraints imposed by the MSF. The attack requires no knowledge of the internal parameters, configurations, or implementation details of either the LiDAR localization module or the MSF. By simultaneously perturbing these sensor modalities, the proposed approach aims to maximize the resulting localization error and expose vulnerabilities in the complete localization pipeline.

5.1 Threat Model

5.1.1 Attacker Model. We assume an adversary with physical access to the environment who can perform an HFR attack against the victim LiDAR sensor using a spoofing device. Similar to the threat model described in Section 3, the attacker injects forged LiDAR pulse reflections to perturb the observation space and influence the LiDAR localization module.

In addition, we assume that the adversary has GNSS spoofing capabilities. Specifically, the attacker can observe the live trajectory of the victim vehicle and transmit carefully crafted counterfeit

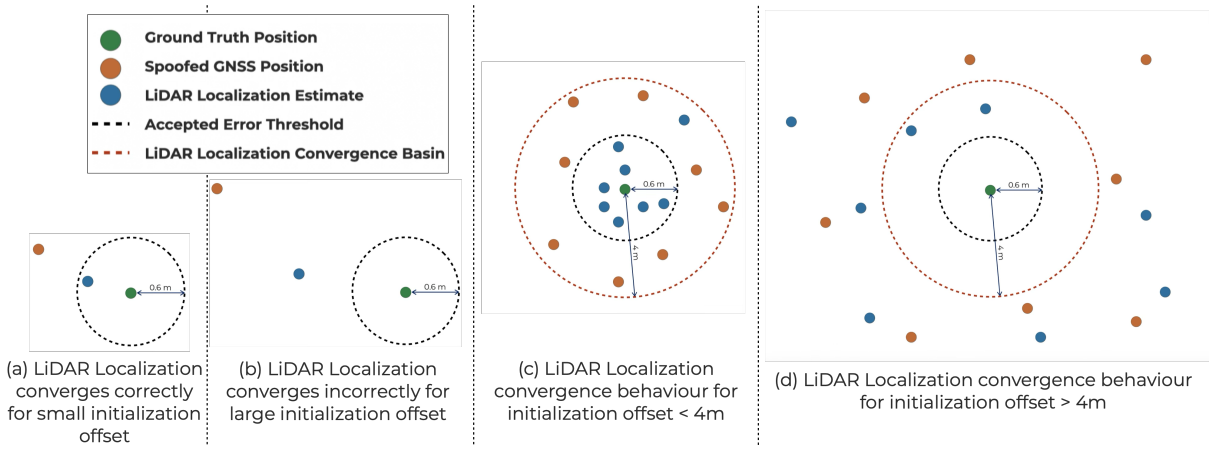


Figure 6: Illustration of LiDAR localization convergence under different GNSS initialization offsets. The black dashed circle represents the 0.6 m accepted localization error threshold, while the red dashed circle represents the empirically determined ~4 m convergence basin obtained through random search. (a) NDT converges correctly when the initialization lies within the convergence basin and near the ground truth. (b) An initialization outside the convergence basin can cause NDT to converge to an incorrect local optimum. (c) Multiple initializations within the ~4 m basin converge toward the correct solution, illustrating the observed convergence behavior for offsets below 4 m. (d) Initializations beyond the ~4 m basin result in localization estimates that fail to converge to the ground-truth position.

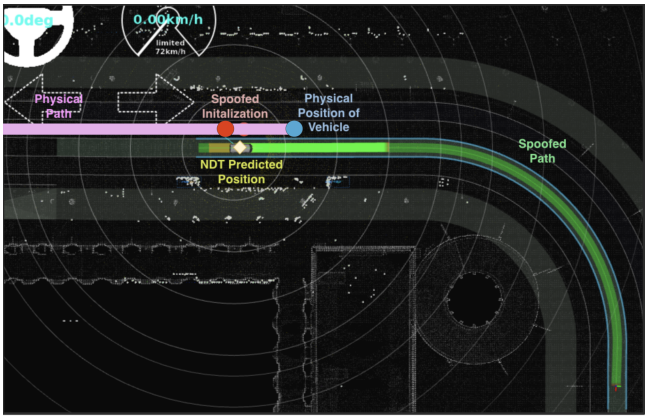


Figure 7: Example of a consequence of Initialization Attack on Autoware. Manipulated initial position biases LiDAR localization towards an incorrect local minimum, causing the navigation system to plan a route from the spoofed position. As a result, the simulated vehicle follows a trajectory opposite to the intended direction and reaches a different destination.

GNSS signals to the victim receiver. This capability follows attacker models commonly adopted in prior GNSS spoofing studies [16, 27], where the objective is to introduce controlled errors into the estimated vehicle position while remaining consistent with the expected vehicle motion.

Unlike isolated attacks that target a single sensing modality, our proposed cross-space attack coordinates LiDAR and GNSS perturbations to amplify localization errors. By observing the victim's live trajectory, the attacker aligns GNSS perturbations with the dominant drift direction of the LiDAR localization. This alignment increases the final localization error while requiring smaller GNSS deviations, thereby reducing the likelihood of triggering trajectory-consistency or anomaly detection mechanisms. Since practical MSF implementations fuse sensor measurements asynchronously using filters such as Kalman filters [17, 29], the perturbation of multiple sensors do not require precise temporal synchronization, making the attack more feasible under realistic operating conditions.

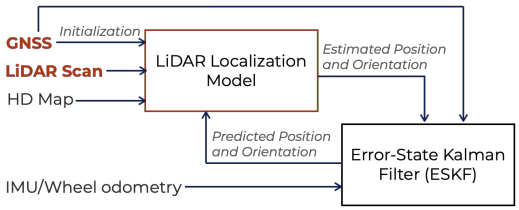


Figure 8: Threat Model for Cross-Space Adversarial Attack. The adversary simultaneously targets the observation space through LiDAR HFR and initialization space through GNSS spoofing. The victim ESKF-based localization MSF fuses GNSS, IMU, and LiDAR localization outputs to estimate the vehicle position.

5.1.2 Victim Model. We consider an Error-State Kalman Filter (ESKF)-based MSF as the victim localization system. Such architectures are widely adopted in autonomous driving platforms, including Baidu Apollo [2], and have been extensively studied in prior security research [16, 27, 29]. As illustrated in Figure 8, the MSF framework fuses GNSS, IMU, and LiDAR localization measurements to estimate the vehicle state. Unlike systems that use GNSS only

for initialization (such as Autoware), the considered ESKF-based MSF framework incorporates GNSS measurements throughout the fusion process.

Our attack targets the interaction among these sensing modalities. In particular, directly introducing large GNSS deviations may trigger anomaly detection mechanisms based on historical trajectories or cross-sensor consistency. We therefore develop an optimization-based attack that leverages the error characteristics observed in observation-space attacks to identify effective perturbations across multiple attack surfaces. The proposed approach jointly optimizes LiDAR observation perturbations and GNSS-induced localization errors to maximize the deviation of the fused localization estimate while satisfying predefined attack constraints.

The victim pipeline incorporates the LiDAR localization modules evaluated in previous sections, A-LOAM and HDL-Localizer, allowing us to evaluate how coordinated perturbations to LiDAR observations and GNSS measurements affect the final fused localization estimate.

5.2 Attack Strategy

The analyses in Section 3 show that, although LiDAR localization errors vary in magnitude and direction across individual frames, their magnitude and direction remain correlated over consecutive frames. This behavior arises because LiDAR localization is an optimization-based inference process rather than a direct positional measurement. We exploit this temporal correlation of LiDAR errors to design a cross-space attack that aligns LiDAR and GNSS errors in the same direction, causing them to accumulate.

As illustrated in Figure 9, we use the previous N errors between spoofed LiDAR localization estimates and their corresponding legitimate GNSS positions to predict the magnitude and direction of the current LiDAR localization error. We then align the GNSS spoofing offset with the predicted error direction to maximize the overall localization error.

We focus exclusively on the *position* component of the localization error and perturb only the horizontal coordinates (x, y), while keeping the vertical coordinate z unchanged. This is consistent with practical autonomous driving systems, where GNSS measurements typically have greater uncertainty in the vertical direction and localization pipelines primarily rely on horizontal position for navigation and planning.

Let the legitimate GNSS position at time step t be

$$\mathbf{g}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix} \tag{3}$$

and the LiDAR localization estimate under HFR attack be

$$\mathbf{l}_t = \begin{bmatrix} x_t^L \\ y_t^L \end{bmatrix}. \tag{4}$$

The instantaneous localization error induced by LiDAR spoofing is then computed with respect to the corresponding legitimate GNSS measurement as

$$\mathbf{e}_t = \mathbf{l}_t - \mathbf{g}_t = \begin{bmatrix} e_t^x \\ e_t^y \end{bmatrix}. \tag{5}$$

The corresponding error magnitude and direction are given by

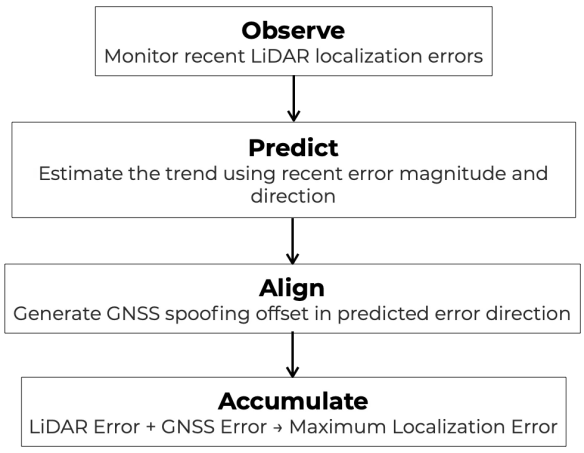


Figure 9: Overview of the proposed Cross-Space Attack. Recent LiDAR localization errors are used to predict the error direction, which guides the GNSS spoofing offset to accumulate LiDAR and GNSS errors and maximize localization error.

$$m_t = \|\mathbf{e}_t\|_2, \tag{6}$$

and

$$\theta_t = \operatorname{atan2}(e_t^y, e_t^x), \tag{7}$$

respectively.

Instead of directly using the instantaneous error, we estimate a stable drift trend using a sliding temporal window. For each time step t, we consider the previous n localization errors,

$$\mathcal{W}_t = \{(m_i, \theta_i) \mid i = t-n, \ldots, t-1\}. \tag{8}$$

Localization errors produced by LiDAR spoofing may occasionally contain outliers arising from transient optimization behavior. Since practical localization MSFs employ statistical outlier rejection mechanisms (e.g., Mahalanobis gating) [29] to discard inconsistent sensor measurements before fusion, we similarly reject outliers during the computation of the representative drift statistics. Specifically, samples whose magnitude deviates significantly from the window mean or whose direction differs substantially from the dominant drift direction are discarded. The remaining observations are used to estimate the representative error magnitude and direction,

$$(\bar{m}_t, \bar{\theta}_t). \tag{9}$$

Because the localization drift direction is an angular quantity, the representative direction is computed using the circular mean,

$$\bar{\theta}_t = \operatorname{atan2}\left(\frac{1}{N}\sum_{i=1}^{N}\sin\theta_i,\ \frac{1}{N}\sum_{i=1}^{N}\cos\theta_i\right), \tag{10}$$

where N denotes the number of retained observations after outlier rejection.

To further increase the induced localization error, we introduce an attacker-controlled offset d to the estimated error magnitude,

$$m_t^* = \bar{m}_t + d. \tag{11}$$

Finally, the spoofed GNSS position is generated by translating the legitimate GNSS measurement along the dominant localization drift direction,

$$\tilde{x}_t = x_t + m_t^* \cos \bar{\theta}_t, \tag{12}$$

$$\tilde{y}_t = y_t + m_t^* \sin \bar{\theta}_t. \tag{13}$$

The resulting spoofed GNSS trajectory is therefore

$$\tilde{\mathbf{g}}_t = \begin{bmatrix} \tilde{x}_t \\ \tilde{y}_t \end{bmatrix}. \tag{14}$$

6 Experimental Evaluation

6.1 Attack Parameters

We evaluate the effectiveness of our proposed cross-space adversarial attack on sequences from the KITTI dataset. We evaluate it against the Error-State Kalman Filter (ESKF) implementation provided by the MATLAB Autonomous Driving Toolbox ('insfilterErrorState') [22].

The attack parameters are selected based on the observations from our previous analysis of observation-space perturbations. Specifically, we set the temporal window size to $n = 5$. This choice is motivated by the error directionality analysis in Table 2, where we observed that a window size of five frames provides a stable estimate of the localization drift while maintaining a low standard deviation in both error magnitude and direction. Therefore, we use this window size to estimate the dominant drift characteristics for generating GNSS perturbations.

To control the magnitude of the GNSS perturbation, we vary the attacker-controlled spoofing offset (d) from 0.2 m to 1.2 m. These perturbation magnitudes are consistent with the range of GNSS spoofing errors considered in prior attacks [16, 27] on localization MSF.

6.2 Results

Table 4 compares the relative positional error achieved by the proposed cross-space adversarial attack with that of observation-space perturbations alone. The results demonstrate that jointly manipulating LiDAR observations and GNSS measurements significantly amplifies the localization error. For A-LOAM, the mean positional error increases by approximately 7 m, while for HDL-Localizer, the increase exceeds 15 m. These results confirm that jointly exploiting the observation and initialization spaces is substantially more effective than perturbing LiDAR observations alone.

Figure 10 illustrates the spoofed LiDAR and GNSS trajectories, together with the resulting fused localization output for A-LOAM under the proposed cross-space attack with a GNSS spoofing offset of d=0.2 m.

We further observe that the relative increase in localization error is larger for HDL-Localizer than for A-LOAM. This difference can be explained by the underlying localization architectures. A-LOAM performs scan-to-scan registration, where the current pose estimate is obtained by aligning consecutive LiDAR scans. Although susceptible to accumulated drift, the temporal constraints imposed by consecutive scan alignment partially mitigate the impact of initialization perturbations. In contrast, HDL-Localizer performs scan-to-map registration using NDT, where the optimizer directly aligns the current scan with a global map. Consequently, an incorrect initialization can move the optimizer outside the basin of attraction of the correct solution, causing convergence to an incorrect local optimum and resulting in larger and more persistent localization errors.

Model	Offset d (m)	Mean (m)	Max (m)
A-LOAM [28]	0.2	6.976	15.898
	0.4	7.172	15.770
	0.6	7.174	15.933
	0.8	7.234	16.171
	1.0	7.388	15.792
	1.2	7.408	15.917
HDL Localizer [11]	0.2	15.288	44.780
	0.4	15.309	44.477
	0.6	15.341	44.607
	0.8	15.355	44.488
	1.0	15.391	44.629
	1.2	15.410	44.621

Table 4: Relative increase in Position Error after Cross-Space Adversarial Attack for varying GNSS Spoofing offsets (d).

Figure 10: Trajectory deviation induced by the cross-space adversarial attack on A-LOAM with GNSS spoofing offset $d = 0.2$ m.

We also observe the effect of varying the GNSS spoofing offset used in the cross-space attack. Across the evaluated range, different spoofing offsets produce comparable localization errors, indicating that the effectiveness of the attack is determined primarily by the direction of the perturbation rather than its magnitude.

Overall, these results demonstrate that the proposed cross-space adversarial attack exploits the interaction between the observation and initialization spaces to achieve substantially greater localization degradation than attacks targeting either component in isolation. The findings further emphasize that securing individual sensors independently is insufficient, as the coupling between GNSS initialization and LiDAR-based localization creates a significantly more powerful attack surface for multi-sensor autonomous driving systems.

7 Discussion

Impact of Historical Position Logs on Initialization Attacks. A potential defense against initialization attacks is to maintain historical vehicle position logs and verify the consistency of the current localization estimate with previously observed trajectories. However, in systems such as Autoware, the final position estimate from a previous trip is obtained relative to the map through fusion of LiDAR localization and IMU measurements. Since this estimate lacks an independent global reference, such as a trusted GNSS measurement, storing historical poses alone provides limited protection against initialization attacks. Effective detection would require maintaining trusted globally referenced position information in addition to local trajectory estimates.

Leveraging Initialization Attack after Localization Failure. During our LiDAR spoofing experiments, we observed a case where HDL-Localizer lost consistency and converged to an incorrect local minimum (shown in Figure 4). Once the localization system enters such a state, initialization attack alone is sufficient to further manipulate the estimated trajectory. By supplying carefully crafted position priors, an adversary can guide the localization process toward a desired incorrect trajectory. However, a practical constraint is that the manipulated trajectory should maintain realistic motion patterns and remain consistent with IMU measurements, similar to the constraints discussed in prior work [14].

Extending Attacks Beyond Position Errors. This work primarily focuses on position errors in localization estimates. However, incorporating orientation perturbations represents an important direction for strengthening cross-space attacks. Orientation errors directly affect vehicle heading and can lead to severe consequences even with small position deviations. For example, an attacker inducing a heading error near an intersection could cause the localization system to estimate an incorrect vehicle orientation, resulting in an unintended maneuver by the planning module. Future work will investigate joint optimization of position and orientation perturbations to develop stronger cross-space attacks and evaluate their impact on autonomous driving decisions.

8 Conclusion

In this paper, we identified the initialization space as a previously unexplored attack surface in ML-based LiDAR localization systems. We demonstrated that manipulating the position prior through GNSS spoofing alone can induce significant localization errors without modifying LiDAR observations. We further analyzed the characteristics of observation-space attacks and showed that the induced localization drift exhibits consistent directional behavior, enabling the formulation of a cross-space adversarial attack that jointly exploits initialization and observation spaces. Our experimental results demonstrate that coordinated perturbations substantially amplify localization errors compared to LiDAR-only attacks while requiring only small GNSS perturbations. These findings highlight the importance of securing not only individual sensors but also the optimization process and cross-modal dependencies within multi-sensor localization frameworks.

Acknowledgments

This project is funded by German Federal Ministry for Digital Affairs and Transport (BMDV) under the project Cypher-AV (Funding Code: 45AVF5A011).

References

[1] Eren Allak, Roland Jung, and Stephan Weiss. 2019. Covariance pre-integration for delayed measurements in multi-sensor fusion. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6642–6649.

[2] Baidu Apollo team. 2017. Apollo: Open Source Autonomous Driving. https://github.com/ApolloAuto/apollo. Accessed: 2026-07-23.

[3] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, 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.

[4] CARLA Simulator Team. 2026. CARLA: Open-source Simulator for Autonomous Driving Research. https://github.com/carla-simulator/carla. Accessed: 2026-07-23.

[5] Pierre Dellenbach, Jean-Emmanuel Deschaud, Bastien Jacquet, and François Goulette. 2022. CT-ICP: Real-time elastic LiDAR odometry with loop closure. In *2022 international conference on robotics and automation (ICRA)*. IEEE, 5580–5586.

[6] 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.

[7] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. 2021. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9710–9719.

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3354–3361.

[9] Nikhil Gosala, Andreas Bühler, Manish Prajapat, Claas Ehmke, Mehak Gupta, Ramya Sivanesan, Abel Gawel, Mark Pfeiffer, Mathias Bürki, Inkyu Sa, et al. 2019. Redundant perception and state estimation for reliable autonomous racing. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 6561–6567.

[10] GPS. 2026. Pseudorandom Noise Code Assignments. https://www.gps.gov/pseudorandom-noise-code-assignments. Accessed: September 15, 2026.

[11] Kenji Koide, Jun Miura, and Emanuele Menegatti. 2019. A portable three-dimensional LIDAR-based system for long-term and wide-area people behavior measurement. *International Journal of Advanced Robotic Systems* 16, 2 (2019).

[12] Yizhen Lao, Yu Zhang, Ziting Wang, Chengbo Wang, Yifei Xue, and Wanpeng Shao. 2025. DisorientLiDAR: Physical Attacks on LiDAR-based Localization. *arXiv preprint arXiv:2509.12595* (2025).

[13] Rokuto Nagata, Kenji Koide, Yuki Hayakawa, Ryo Suzuki, Kazuma Ikeda, Ozora Sako, Qi Alfred Chen, Takami Sato, and Kentaro Yoshioka. 2025. Slamspoof: Practical lidar spoofing attacks on localization systems guided by scan matching vulnerability analysis. In *Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 15181–15187.

[14] Sashank Narain, Aanjhan Ranganathan, and Guevara Noubir. 2019. Security of GPS/INS based on-road location tracking systems. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 587–601.

[15] Takami Sato, Yuki Hayakawa, Ryo Suzuki, Yohsuke Shiiki, Kentaro Yoshioka, and Qi Alfred Chen. 2024. LiDAR Spoofing Meets the New-Gen: Capability Improvements, Broken Assumptions, and New Attack Strategies. In *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS)*. ISOC.

[16] Junjie Shen, Jun Yeon Won, Zeyuan Chen, and Qi Alfred Chen. 2020. Drift with devil: Security of {Multi-Sensor} fusion based localization in {High-Level} autonomous driving under {GPS} spoofing. In *Proceedings of the 29th USENIX security symposium (USENIX Security 20)*. 931–948.

[17] Apollo Shenlan. 2017. Shenlan msf. https://github.com/shenlan2017/Apollo_ShenLan. Accessed: 2026.

[18] Héber Sobreira, Carlos M Costa, Ivo Sousa, Luis Rocha, José Lima, PCMA Farias, Paulo Costa, and A Paulo Moreira. 2019. Map-matching algorithms for robot self-localization: a comparison between perfect match, iterative closest point and normal distributions transform. *Journal of Intelligent & Robotic Systems* 93, 3 (2019), 533–546.

[19] Todor Stoyanov, Jari Saarinen, Henrik Andreasson, and Achim J Lilienthal. 2013. Normal distributions transform occupancy map fusion: Simultaneous mapping and tracking in large scale dynamic environments. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 4702–4708.

[20] Yuna Tanaka, Kazuki Nomoto, Ryunosuke Kobayashi, Go Tsuruoka, and Tatsuya Mori. 2025. {WIP}: Understanding the Mechanisms Behind {NDT-Based} Localization Vulnerabilities in Autonomous Driving. In *3rd USENIX Symposium on Vehicle Security and Privacy (VehicleSec 25)*. 211–219.

[21] The Autoware Foundation. 2026. Autoware: Open Source Software for Autonomous Driving. https://github.com/autowarefoundation/autoware/. Accessed: 2026-07-23.

[22] The MathWorks, Inc. 2026. Automated Driving Toolbox. https://www.mathworks.com/products/automated-driving.html

[23] Nils Ole Tippenhauer, Christina Pöpper, Kasper Bonne Rasmussen, and Srdjan Capkun. 2011. On the requirements for successful GPS spoofing attacks. In *Proceedings of the 18th ACM conference on Computer and communications security*. 75–86.

[24] USENIX Association. 2026. USENIX Security '26 Call for Papers: Ethics Guidelines. https://www.usenix.org/conference/usenixsecurity26/call-for-papers#ethics. Accessed: 2026-07-24.

[25] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. 2023. Kiss-icp: In defense of point-to-point icp–simple, accurate, and robust registration if done the right way. *IEEE Robotics and Automation Letters* 8, 2 (2023), 1029–1036.

[26] Guowei Wan, Xiaolong Yang, Renlan Cai, Hao Li, Yao Zhou, Hao Wang, and Shiyu Song. 2018. Robust and precise vehicle localization based on multi-sensor fusion in diverse city scenes. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 4670–4677.

[27] Junqi Zhang, Shaoyin Cheng, Linqing Hu, Jie Zhang, Chengyu Shi, Xingshuo Han, Tianwei Zhang, Yueqiang Cheng, and Weiming Zhang. 2025. The Ghost Navigator: Revisiting the Hidden Vulnerability of Localization in Autonomous Driving. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security 25)*. 3979–3998.

[28] Ji Zhang and Sanjiv Singh. 2014. LOAM: Lidar odometry and mapping in real-time. In *Robotics: Science and systems*, Vol. 2. Berkeley, CA, 1–9.

[29] Qifan Zhang, Junjie Shen, Mingtian Tan, Zhe Zhou, Zhou Li, Qi Alfred Chen, and Haipeng Zhang. 2022. Play the imitation game: Model extraction attack against autonomous driving localization. In *Proceedings of the 38th Annual Computer Security Applications Conference*. 56–70.

[30] Zhe Zhang, Shaoshan Liu, Grace Tsai, Hongbing Hu, Chen-Chi Chu, and Feng Zheng. 2018. Pirvs: An advanced visual-inertial slam system with flexible sensor fusion and hardware co-design. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3826–3832.

A Open Science

To facilitate reproducibility and allow the research community to verify our methodology, we provide an artifact package containing the necessary source code, dataset files, and scripts used in our evaluation:

- **Artifact Contents:**
  - **Attack Scripts:** Code for synthesizing GNSS-induced position prior perturbations and conducting simultaneous GNSS and LiDAR spoofing attack.
  - **Evaluation Pipelines:** Dataset files and wrapper scripts for integrating and evaluating victim MSF inside MATLAB environment.
  - **Data Processing & Plotting:** Analysis scripts used to calculate trajectory drift, directional bias, and convergence boundaries across KITTI dataset sequences.
- **Artifact Access:** All the scripts, datasets, and reproduction steps are accessible at: https://github.com/pranavshetty17/Cross-Space-Attacks
- **Restrictions & Disclosure Safeguards:** Full replication of our experiments requires standard tools (Python, MATLAB, Autoware, CARLA simulator, ROS) and publicly accessible benchmarks (KITTI). Hardware specifications for the optical laser spoofing setup are discussed at a conceptual level based on prior literature to prevent premature weaponization of physical spoofing hardware while preserving complete algorithmic reproducibility.

B Ethical Considerations

This research investigates security vulnerabilities in autonomous driving localization pipelines, specifically NDT and MSF systems combining GNSS, LiDAR, and IMU measurements. While our work highlights actionable exploit vectors (initialization and joint cross-space attacks), our goal is to improve the resilience and safety of autonomous vehicles by identifying architectural weaknesses in local optimization dependencies.

To ensure responsible research practices and eliminate deployment risks, our study strictly follows the ethical framework established by the USENIX Security Ethics Policy [24]:

- **Simulation and Dataset Constraints:** All experimental evaluations were conducted strictly within controlled software simulation environments (CARLA operating with Autoware) and publicly available benchmarks (the KITTI dataset). No experiments were executed on physical vehicles or public roads, eliminating any risk to public safety, physical infrastructure, or human participants.
- **Dual-Use Balance:** The attack methodologies described in this paper serve to demonstrate fundamental limits of local optimizers within multi-sensor architectures. We believe the defensive benefit of understanding cross-modal vulnerabilities and multi-sensor dependencies significantly outweighs the risks, as it enables the development of robust initialization validation and multi-sensor consistency checks.

C Use of Generative AI Tools

Generative AI tools (such as ChatGPT, Gemini, and Claude) were employed during the preparation of this work to assist with text regeneration, grammar refinement, and code generation. Specifically, these tools were used to edit and polish human-written text for conciseness and stylistic clarity, as well as to assist in writing implementation scripts based on explicit algorithmic descriptions provided by the authors. All generated code and text were thoroughly reviewed, manually verified, and tested by the authors, who retain full responsibility for the contents and integrity of this paper.