# ADoPT: LiDAR Spoofing Attack Detection Based on Point-Level Temporal Consistency

Minkyoung Cho[1]
minkycho@umich.edu
Yulong Cao[2]
yulongc@nvidia.com
Zixiang Zhou[1]
zixiangz@umich.edu
Z. Morley Mao [1]
zmao@umich.edu

[1] Computer Science & Engineering
University of Michigan
Ann Arbor, MI, USA

[2] NVIDIA Research
Santa Clara, CA, USA

**Abstract**

Deep neural networks (DNNs) are increasingly integrated into LiDAR (Light Detection and Ranging)-based perception systems for autonomous vehicles (AVs), requiring robust performance under adversarial conditions. One pressing concern is the challenge posed by LiDAR spoofing attacks, where attackers inject fake objects into LiDAR data, leading AVs to misinterpret their surroundings and make faulty decisions. Many current defense algorithms predominantly depend on perception outputs, such as bounding boxes. However, these outputs are intrinsically limited as they are generated by imperfect perception models that process a restricted set of points, acquired from the ego vehicle's specific viewpoint. The reliance on bounding boxes is a manifestation of this fundamental constraint. To overcome these limitations, we propose a novel framework, named ADoPT (**A**nomaly **D**etection based **o**n **P**oint-level **T**emporal consistency), which quantitatively measures temporal consistency across consecutive frames and identifies abnormal objects based on the coherency of point clusters. In our evaluation using the nuScenes dataset, our algorithm effectively counters various LiDAR spoofing attacks, achieving a low (< 10%) false positive ratio and high (> 85%) true positive ratio, outperforming existing state-of-the-art defense methods, CARLO and 3D-TC2. Moreover, ADoPT shows promising potential for accurate defense in diverse road environments.

## 1 Introduction

The growing incorporation of deep neural networks (DNNs) in LiDAR (Light Detection and Ranging)-based perception for autonomous vehicles (AVs) calls for rigorous attention to their robust performance. In light of this challenge, researchers are focused on developing and refining various defense technologies for potential attacks targeting AV perception systems. One prominent research direction in this field involves the manipulation of LiDAR point cloud data. Attackers can fabricate data by jamming and relaying original LiDAR signals [19, 32], emitting spurious LiDAR purses [13, 21, 22, 23], or exploiting vulnerabilities in the DNN-based perception module [24, 25, 34], causing AVs to misinterpret their driving environment and make faulty decisions (e.g., emergency alarm activation, sudden breaking, lane changing, etc).

Defense algorithms based on AV perception outputs (i.e., bounding boxes) have been widely studied. Figure 1 showcases two state-of-the-art bounding box-based algorithms: (1) physical principles-based approach (e.g. CARLO) [7, 23, 31], which detects attacks by leveraging physical principles governing authentic objects; and (2) temporal consistency-based method (e.g. 3D-TC2) [16, 30, 35], which focuses on motion consistency across adjacent frames. While temporal consistency offers an edge over physical principles, both are fundamentally limited by their dependence on bounding boxes. Given the constraints of an ego vehicle's viewpoint and the inherent inaccuracies of perception modules (especially for small and distant objects) [15, 36, 37], relying on bounding boxes proves inadequate and leads to inaccuracies in anomaly detection.

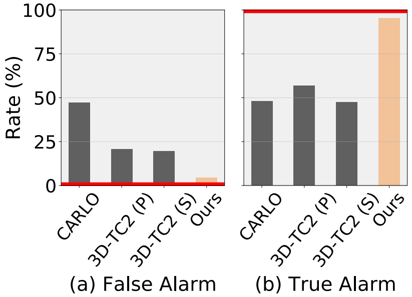


Figure 1: Our ADoPT method outperforms existing methods [23, 35] with a 4.4 ∼ 10.5× lower false alarm rate and 1.7 ∼ 2× higher true alarm rate. The ideal cases are represented by the red solid lines with a false alarm rate of 0 and a true alarm rate of 100. We used PointPillars [9] for CARLO and 3D-TC2 (P) and SECOND [33] for 3D-TC2 (S) for 3D object detection. False alarms arise when benign LiDAR frames are misidentified as attacked, while true alarms occur when poisoned frames, created by injecting simulated pedestrian points, are correctly recognized as attacked. These results show existing defenses often output a bounding box that does not tightly fit the object or is false and struggle to detect small fake objects.

We introduce ADoPT (**A**nomaly **D**etection based **o**n **P**oint-level **T**emporal consistency) to build perception model-agnostic monitoring modules. Harnessing the rich and comprehensive information present in raw sensor data [4, 36], our approach offers profound advantages in defending against LiDAR spoofing attacks [23, 31, 35], bypassing the limitations of traditional perception models. While raw sensor data provides extensive information, implementing defense algorithms (e.g. temporal consistency) based on raw sensor data is challenging. Our approach emerges from our observation of the intuitive notion that an object consists of point clusters with a specific degree of point intensity, moving coherently. This understanding enables the measurement of temporal consistency at the point cloud level. Utilizing our temporal consistency foundation, we present a two-stage approach to detect adversarial manipulations in frames. Initially, our coherence-enhanced scene flow estimation predicts expected object locations while maintaining coherency, even in the face of point injections, outperforming conventional methods susceptible to anomalies. This robust estimation sets a firm foundation for the subsequent phase: clustering-based anomaly detection. It contrasts point clusters between expected and observed point locations, with discrepancies between them highlighting potential adversarial interventions.

ADoPT stands robust against both dense and sparse point injection LiDAR spoofing attacks, achieving a commendable false positive ratio (FPR) of less than 10% and a true positive ratio (TPR) exceeding 85%, thereby surpassing existing defense mechanisms grounded on perception output. Additionally, we highlight the effectiveness of our anomaly detection metric based on cluster coherency, showcasing its superiority over traditional methods through comparative analysis.

# 2 Related Work

Autonomous vehicle (AV) defense mechanisms against sensor data fabrication attacks are broadly classified into physical principle-based and consistency-based methods.

**Physical Principle-based Defense.** These techniques leverage specific geometries or physical-invariant properties, which attackers struggle to imitate when forging objects. CARLO [23] employs free/occluded space or laser rays within a frustum space related to each detected bounding box to distinguish between real and fake objects. Shadow-Catcher [7] utilizes shadow region differences based on bounding box coordinates, and LOP [31] introduces the concept of *objectness* by considering the point density and the distance from the LiDAR sensor to the predicted objects. These approaches use specialized rules to assess an object's adherence to physical principles.

**Consistency-based Defense.** These methods emphasize temporal consistency and show promising detection success rates for AV systems. They exploit the invariant nature of object motion across consecutive frames. AdvIT [30] counters adversarial attacks on video frames, where the attacker manipulates the distribution of points but preserves the appearance of the original points, by estimating the optical flow of each pixel and measuring temporal consistency. PercepGuard [16] utilizes spatio-temporal consistency for misclassification attacks, where the attacker alters the labels of detected outputs on camera images (e.g., from car to people), and verifies moving patterns of bounding boxes. 3D-TC2 [35] proposes a temporal consistency check-based method to detect LiDAR spoofing attacks, converting LiDAR point clouds into 2D images and comparing predicted motion to detected bounding boxes.

**Limitations.** Existing studies rely on perception modules, assuming their high accuracy. However, raw sensor data processing using perception modules can result in false detections or information loss, especially for small objects like pedestrians and cyclists – critical objects that autonomous vehicles must consider in making navigation decisions. Our work introduces a novel paradigm for attack detection algorithms, using only raw sensor data to achieve robust defense regardless of the object type. We detail the core components enabling point-level anomaly detection in the following sections.

# 3 Background: Scene Flow Estimation

In the 3D point cloud domain, scene flow represents the 3D motion of each point across consecutive frames. Accurate scene flow estimation is crucial for predicting user or AV motion and estimating trajectories. However, real-world estimation remains challenging due to temporal occlusion and dynamic, rigid object motions. Scene flow estimation has evolved into two primary branches.

**Offline Learning Methods.** These approaches [14] use separate offline training processes with annotated datasets. Scene flow estimation is formulated as a DNN model that receives a pair of frames and outputs the optimal flow. DNN models offer customizability and high capacity for flow representation, achieving high accuracy while addressing bad correspondences. However, they face limitations such as requiring substantial data and ground truth labels, which are difficult to obtain [2, 3]. Researchers generate labels using alternative methods [20] or employ self-supervised learning [17, 29]. These methods may struggle with input frames deviating from the training dataset due to low generalization capabilities [11].

**Online Optimization Methods.** These approaches do not require separate training processes or datasets, instead formulating scene flow estimation as an optimization problem. They demonstrate higher accuracy on out-of-distribution point cloud frames, which fall outside the training dataset used by offline learning-based methods, making them more suitable for real-world situations.

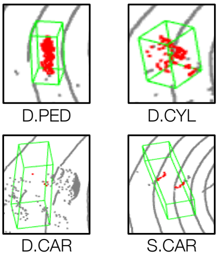


Figure 2: Examples of Injected Points. D.PED, D.CYL, and D.CAR are fake objects created by dense point injection attacks, mimicking a pedestrian, cyclist, and car, respectively. S.CAR results from a sparse point injection attack imitating a car. AVs recognize these objects as real (green bounding box). Red points are within the bounding box, while gray points are outside, suggesting that bounding boxes may not always adequately fit objects.

Classic methods, such as Iterative Closest Point (ICP) [1], initialize flow vectors for the point within the point cloud frame, solving optimization problems to find the optimal flow vectors that represent the discrepancy of two input frames at runtime. Recent studies have proposed various solutions with high generalizability by formulating scene flow estimation in diverse ways, such as multi-layer perceptron (MLP) [11, 12], graph Laplacian [20], or Bayesian inference [8]. Among these methodologies, NSFP [11] introduced an advanced method by changing flow representation from displacement vectors to an MLP-based model. With this MLP-formed flow, they solve the optimization problem and find the optimal flow at runtime while iteratively adjusting the MLP parameters like DNN training.

# 4 Threat Model

We investigate two spoofing attacks: dense and sparse point injection. Dense point injection attacks inject up to 200 points and achieve a high Attack Success Rate (ASR) of 96%-97%, producing a visually recognizable fake object. 3D-TC2 [35] is designed to counter this attack and serves as our evaluation baseline. Conversely, sparse point injection attacks [23] inject up to 64 points, rendering the fake object difficult to visually identify, with an ASR of less than 21%. CARLO [23] is a proposed defense method to combat this sparse injection attack, used as our evaluation baseline. Figure 2 displays examples of spoofed objects.

# 5 ADoPT Methodology

In this section, we introduce ADoPT, a solution for point-level anomaly detection devised to enhance the resilience of object detection systems. Leveraging the observation that injected points demonstrate *poor temporal consistency* — appearing inconsistently within the point cloud frame over time — ADoPT utilizes scene flow estimation to quantify objects' temporal consistency, thereby facilitating the detection of point injection attacks. Figure 3 illustrates overall ADoPT architecture, where $F_1, F_2, ..., F_L$ are sequential historical LiDAR point cloud frames, and $F_{L+1}$ is the subsequent incoming frame. Initially, ADoPT generates a synthesis of preceding frames by aligning all points from the historical frames using scene flow estimation (Sec. 5.1). This synthesized representation is then compared with the incoming frame to identify points that showcase inadequate temporal alignment, earmarking them as potential injections from attackers (Sec. 5.2). Furthermore, we outline several techniques to mitigate runtime overhead, thereby making ADoPT a viable solution when implemented in AV systems (Sec. 5.3).

## 5.1 Coherence-Enhanced Scene Flow Estimation

For quantification of temporal consistency, central to ADoPT is the process of aligning points captured at different timestamps and combining them into a single frame. Scene flow estimation (SFE) is crucial for aligning point cloud frames by calculating optimal point displacement [1, 10]. For generalizability to various road environments and different injected

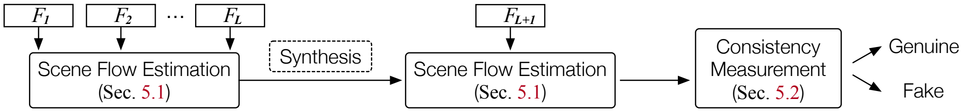

$F_1$
$F_2$
···
$F_L$
Scene Flow Estimation (Sec. 5.1)
Synthesis
$F_{L+1}$
Scene Flow Estimation (Sec. 5.1)
Consistency Measurement (Sec. 5.2)
Genuine
Fake

Figure 3: ADoPT Architecture.

objects, we employ an MLP-formed neural prior to represent scene flow and optimize the MLP parameters at runtime for a pair of point cloud frames ($F_1$, $F_2$), inspired by NSFP [11].

SFE serves a dual purpose: it aligns points in historical frames and juxtaposes the resulting synthesized frame with newly arriving one. However, ensuring precise scene flow becomes challenging without confirmed temporal consistency. In the presence of a LiDAR spoofing attack, conventional methods frequently falter, predicated on the assumption of consistent object appearances across frames. This vulnerability arises because this assumption can be violated through the continued use of SFE over a series of frames, or through the introduction of spurious objects. Repeatedly deploying SFE may amplify errors, leading to a dispersion effect in the synthesis. Additionally, the presence of fabricated points in recent frames hampers the precise functioning of SFE, creating erroneous point correspondences between synthesized and fake points.

To counteract these challenges, we have conceptualized temporal consistency by viewing objects as cohesive point clusters with inherent intensity. Recognizing this coherence, we introduce a loss function that enhances the coherence between the motion flows of points belonging to each cluster (i.e., part of an object). Here, we use a clustering method, DBSCAN [5], a well-established spatial clustering algorithm, to verify if two given points are part of the same cluster. By defining the following loss term, we can enforce neighboring points to move coherently:

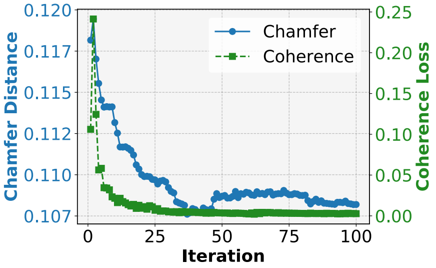


Figure 4: As the number of iterations increases, the two loss values converge harmoniously without impeding each other's individual convergence.

$$L_{coherence}(F_1) = \frac{1}{N^2} \sum_{p_i,p_j \in F_1} (M(p_i,p_j) \cdot w(p_i,p_j) \cdot ||fl(p_i) - fl(p_j)||^2)$$

where $M(p_i, p_j)$ is a binary clustering mask indicating whether any two points in $F_1$, $p_i$ and $p_j$, belong to the same cluster. Concurrently, $w(p_i, p_j)$ is the weight value between these points, formulated to foster more coherent movement between closer points by being influenced by the distance between them. $fl(p_i)$ is a 3-dimensional flow vector that indicates the displacement of point $p_i$ along the x, y, and z axes. $N$ denotes the number of points included in valid clusters (i.e., non-outlier points) as identified through DBSCAN.

Consequently, our final loss function is defined by a combination of the Chamfer Distance (CD) [6] and coherence loss, enabling us to find the optimal scene flow that represents the point motions between $F_1$ and $F_2$ while preventing any point from deviating from its original cluster.

$$L = \alpha L_{chamfer}(F_1,F_2) + \beta L_{coherence}(F_1)$$

Here, $L_{chamfer}(F_1,F_2)$ is the CD value, the most popular distance metric for two point cloud sets, which is defined as:

$$L_{chamfer}(F_1,F_2) = \sum_{p_i \in F_1} \min_{q_j \in F_2} ||p_i - q_j||_2^2 + \sum_{q_j \in F_2} \min_{p_i \in F_1} ||p_i - q_j||_2^2$$

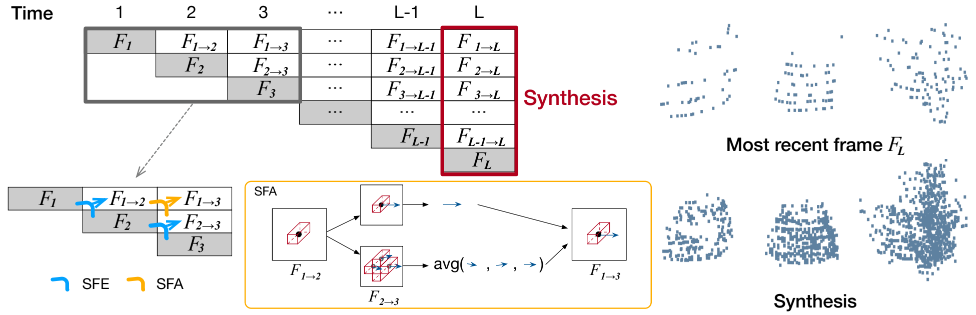


Figure 5: Synthesis Generation Procedure. Given historical frame length $L$, a new LiDAR frame enters the system for Scene Flow Estimation (SFE). The estimated flow propagates through Scene Flow Approximation (SFA) to upper cells. The top right figure shows the actual input frame $F_L$, and the bottom right displays the generated synthesis with clear and dense object shapes (three cars, the rightmost one is obscured behind a tree).

where $p_i$ and $q_j$ represent individual points in $F_1$ and $F_2$ respectively. Figure 4 showcases the functioning of our proposed loss function during the online optimization process, emphasizing the harmonious convergence of the two loss terms.

## 5.2 Cluster-based Consistency Measurement

Implementing the proposed SFE method produces a warped synthesis by predicting the appearance of individual points at the moment each incoming frame arrives. A pivotal stage in ADoPT is the temporal consistency measurement between the resulting synthesis and the incoming frame. However, results obtained using conventional distance metrics (e.g., Chamfer Distance), often lead to substantial variances, thereby making the discrimination process markedly challenging. This fluctuation is predominantly driven by the differing number of points in individual frames, a variable greatly affected by the intricacies of various road scenarios. Moreover, the unpredictable changes introduced by attackers, affecting both the location and the number of points, deem traditional handcrafted schemes inadequate. This necessitates a suitably designed metric.

Our approach employs a cluster-based metric based on the understanding that objects naturally form groups of point clusters. Recognizing the potential in the overlapping characteristic of the synthesis and incoming frame provides a strategy for assessing their consistency. To foster this strategy, we first meld the warped synthesis with the incoming frame, distinctly marking each point to denote whether it originates from the synthesis or the incoming frame. This meticulous categorization aids in identifying areas of inadequate temporal alignment, enabling reliable detection of fabricated objects. Subsequently, ADoPT uses DBSCAN to identify local clusters, effectively discarding outliers, and then removes clusters that contain synthesis points, a process which verifies the presence of genuine objects consistent with historical objects. In benign scenarios, this tactic results in a complete absence of clusters, a testament to the adequate temporal alignment across frames. However, in poisoned scenarios, clusters exclusively formed of incoming frame points remain, serving as indicators of fabricated objects. Illustrative examples are shown in Appendix A.

## 5.3 Additional Components for Runtime Overhead Reduction

**Synthesis Generation on Historical Frames.** To achieve our goal of identifying anomalous objects, ADoPT needs to differentiate between suddenly appearing normal objects and maliciously placed objects. ADoPT synthesizes past frames by warping them to the time of the last historical frame $F_L$ using SFE, inspired by point cloud densification techniques [11, 27]. However, these methods are computationally intensive, and generating synthesis may take up to $O(NL^2)$ time, assuming latency in solving optimization problems is $N$, and we use $L$ historical frames. For instance, with a 0.1-second latency and 10 historical frames, it would take 10 seconds to process all the scene flow estimations, which is impractical and cannot be concealed through parallelization.

To reduce computational complexity, ADoPT approximates scene flow instead of solving optimization problems. The estimated scene flow is propagated to its preceding frames, facilitating a single estimation of the scene flow at each timestamp. As LiDAR senses a varying number of points, voxelization is utilized to identify corresponding points in order to propagate the most recent scene flow into past frames. This process involves mapping voxel indices between frames to integrate the corresponding voxel's scene flow with the target point. In instances where there are no mapped points, the target voxel's scene flow is computed by taking the average of the adjacent voxels' scene flows, a strategy depicted in Figure 5. This substantially reduces the time complexity of computation-intensive SFE processes from $L^2$ to $L$, saving time and enabling parallelization with the SFE process.

**Voxel Downsampling.** In ADoPT, the total latency is primarily influenced by the number of points. To optimize the system, minimizing latency without significant TPR loss and FPR gain, we reduce the number of points via voxelization. This approach's efficacy is detailed in Sec. 6.1, where we illustrate the interplay between voxel grid size, total latency, and the accuracy of attack detection. This highlights the critical role of selecting the optimal voxel grid size to balance system accuracy with timely execution.

## 6 Evaluation

In this section, we evaluate ADoPT under different attack scenarios. Additionally, we compare our proposed algorithm with two widely-used defense methods for LiDAR spoofing attacks, CARLO and 3D-TC2. All experiments are conducted on a server equipped with two Intel Xeon 4110 CPUs and one NVIDIA RTX 2080 GPU.

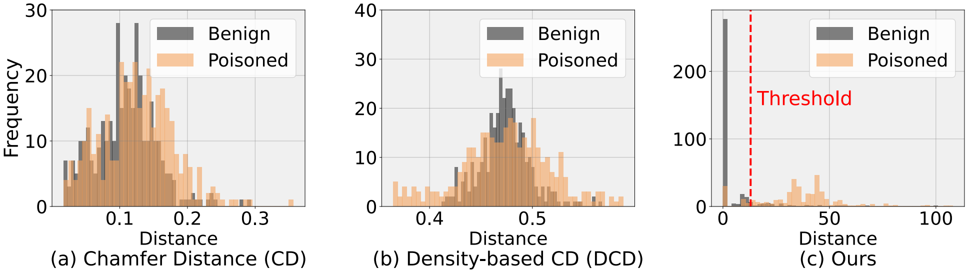
**Dataset.** Our evaluations are performed on the nuScenes dataset [2], a large-scale autonomous driving dataset collected from vehicles equipped with a 32-beam LiDAR system. The nuScenes dataset is divided into two subsets: v1.0-mini (comprising 10 scenes) and v1.0-trainval (comprising 350 scenes). Each scene is 20 seconds long and annotated at a frequency of 2 Hz.

**Attack Scenarios.** The poisoned dataset utilized in the dense point injection attack is sourced from the authors of 3D-TC2, who leveraged the v1.0-mini dataset. In this attack, spoofed data points that represent vehicles, cyclists, and pedestrians are introduced systematically. Concurrently, in addressing the sparse point injection attack, we generated 355 poisoned frames using the validation set of the v1.0-trainval dataset, which consists of 150 scenes, adhering to the approach presented in CARLO. Given the intrinsic sparsity characteristic of this attack, only spoofed points representative of vehicles are introduced.

**Parameter Setting.** We opt for a multi-layer perceptron (MLP) architecture composed of six layers and 128 hidden units, a configuration empirically determined to yield the highest accuracy while maintaining a low latency on our dataset. Figure 4 shows loss convergence after 30 iterations, influencing our choice of a 30-iteration count. For training, we use a fixed learning rate of 0.0008, empirically derived for optimum performance and convergence potential. In defining our loss function, we attribute values of 1 and 2 to variables $\alpha$ and $\beta$, respectively, and the weights between the points, $w(p_i, p_j)$, are all set equally to 1. The DBSCAN procedure relies heavily on two critical thresholds: the minimal distance between

Table 1: Comparison of Defense Methods. A lower false positive rate (FP) and a higher true positive rate (TP) indicate a more accurate attack detection.

| | Dense Point Injection | | | | Sparse Point Injection | |
|---|---|---|---|---|---|---|
| | FP ↓ | TP (D.CAR) ↑ | TP (D.CYL) ↑ | TP (D.PED) ↑ | FP ↓ | TP (S.CAR) ↑ |
| CARLO [23] | 47.2 | 48.0 | 49.4 | 48.0 | 47.9 | 54.4 |
| 3D-TC2 (PP) [35] | 20.7 | 98.6 | 95.0 | 56.9 | 16.6 | 53.5 |
| 3D-TC2 (SEC) [35] | 19.6 | 98.3 | 45.8 | 47.5 | 16.3 | 84.2 |
| ADoPT | 4.5 | 97.2 | 98.3 | 95.2 | 9.3 | 85.4 |



(a) Chamfer Distance (CD) (b) Density-based CD (DCD) (c) Ours

Figure 6: Comparison of Anomaly Detection Metrics: (a) Chamfer Distance, (b) Density-based Chamfer Distance, and (c) Our Proposed Cluster-based Metric. The x-axis represents the distance values, and the y-axis indicates the number of frames corresponding to each value. Our approach allows for establishing a threshold for attack detection, unlike conventional metrics. The difference between the average distance values for benign and poisoned cases supports this claim: (a) benign: 0.11, poisoned: 0.13, (b) benign: 0.47, poisoned: 0.47, (c) benign: 4.31, poisoned: 35.34. In this work, we set our threshold to 15.

the nearest points and the minimal count of points necessary to form a valid cluster. Carefully optimizing for FPR and TPR, we established thresholds of at least 17 points and a 0.25 distance parameter for dense point attacks, and a minimum of 9 points with a 0.75 distance threshold for sparse point attacks. Detailed insights into the threshold determination process are elaborated in Appendix B.

## 6.1 Experimental Results

**Effect of ADoPT.** To substantiate the effectiveness of ADoPT, we benchmark its performance against baseline methods under dense and sparse point injection attack scenarios (see Table 1). In this evaluation, we focus on two pivotal metrics: the false positive rate (FPR), denoting the incorrect identification of benign frames as attacked, and the true positive rate (TPR), reflecting the correct detection of poisoned frames. Our method manifests a low FPR, evincing its efficacy in curtailing false alarms during the detection of spoofing attacks. We match the baseline accuracy on relatively large objects such as D.CAR, while substantially exceeding it when it comes to smaller objects, notably achieving TPRs of 98.3% and 95.2% for cyclists and pedestrians, respectively. These statistics underline ADoPT's superior ability to pinpoint small spoofed objects. Consequently, ADoPT demonstrates marked supremacy in identifying LiDAR spoofing attacks across different object types and attack scenarios.

**Clustering-based Metric for Anomaly Detection.** In considering alternative design approaches for consistency measurement, we acknowledge the potential utility of established distance metrics. Among the prevalent metrics for evaluating point cloud similarity are CD and Earth Mover's Distance (EMD) [6]. Despite its utility, EMD's computational demands deem it unfit for attack detection. To this end, our analysis leverages the Density-based Chamfer Distance (DCD) [28], a novel metric that synergizes the strengths of CD and EMD, promising enhanced accuracy.

As illustrated in Figure 6, both CD and DCD exhibit substantial fluctuations according to road configurations, with object count and road environment serving as prominent influencing factors. This variability creates a challenging environment for establishing the ideal thresholds to distinguish between benign and poisoned frames. In contrast, our clustering-based metric enables threshold determination for attack decisions, making it the most reasonable metric for our situations.

**Ablation Study.** We conducted an ablation study to assess the impact of including or excluding the coherence loss term, as well as employing DCD. Further details can be found in Figure 7. Utilizing DCD as our loss function in dense point injection attack scenarios yielded a 4.6% FPR and TPRs of 94.8%, 92.9%, and 94.0% for D.CAR, D.CYL, and D.PED, respectively. These findings, exhibiting stable FPR and decreased TPRs, support the idea that the injected fake points compromise accurate SFE, emphasizing the effectiveness of the coherence-enhanced SFE.

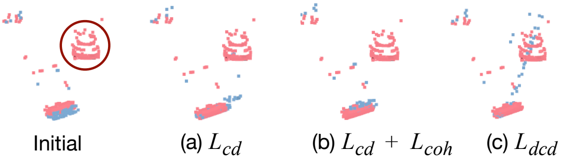

Initial (a) $L_{cd}$ (b) $L_{cd} + L_{coh}$ (c) $L_{dcd}$

Figure 7: Ablations with/without Coherence Loss. We examine the alignment of two point cloud sets (red and blue points) by comparing the warped blue points (obtained by adding the estimated scene flow) to the red points. The first image denotes their initial status, and the car (circled in red) indicates a fake object. (a) shows results using only CD without coherence loss; (b) displays outcomes of our proposed method using both CD and coherence loss. We also explore replacing our loss with DCD in (d). This comparison reveals that fake points hinder SFE, seemingly drawing the red points towards them, highlighting the necessity of our coherence-enhancing loss under adversarial settings.

**Impact of Historical Frame Length.** Historical frames help distinguish sudden appearances of valid and fake points by correlating them with previously observed points. We utilize 10 historical frames at a 10 Hz frequency to align with 3D-TC, allowing for a fair comparison of the effects of using raw data vs. using detection output. Testing with various frame lengths produces nearly consistent results from lengths of 2 to 15, but shows a decline thereafter (Figure 8).

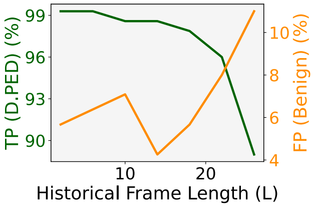


Figure 8: Attack Detection Accuracy Changes in Relation to Historical Length

## 6.2 Limitation & Analysis

**Runtime Overhead.** In ADoPT, the primary contributor to total latency is SFE latency, significantly influenced by the voxel shape. As illustrated in Figure 9, ADoPT currently experiences a considerable latency of 2.1 seconds. Nonetheless, we have the capacity to diminish this to a mere 0.7 seconds, sustaining a robust attack detection accuracy at roughly 87% of the initial accuracy. This indicates a marked enhancement in speed without a critical sacrifice in detection capability. To further hasten this process, we envision incorporating mixed-precision training techniques [18], utilizing a 16-bit (or lower) numerical format for MLP parameters. We also aim to reduce runtime overhead through the pre-training of the MLP model, coupled with the application of test-time training [26], a strategy for quick convergence and adaptability to various incoming frames.

**Failure Cases**. As we employ the spatial clustering method for attack detection, most failure cases arise when spoofed objects are attached to benign road objects. Although classified as false negatives, the spoofed object is identified as part of the benign object it is attached to; thus, it does not significantly affect existing navigation decisions or trigger numerous sudden alarms. Refer to Appendix C for point cloud data illustrating a failure situation.

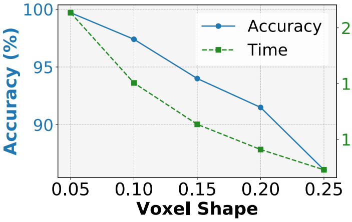


Figure 9: Impact of Voxel Shape on Attack Detection Accuracy and Time. We use a cubic-shaped voxel, and the value of the x-axis represents the length of one side of the voxel.

**Analysis on Impact per Component on Performance.** ADoPT's performance is contingent upon the accuracy of both SFE and DBSCAN. However, thanks to its structural advantage, it is able to maintain effective responsiveness even when there is a decrease in the accuracy of either of these elements. Let's explore this further: Firstly, low SFE accuracy can lead to flawed synthesis and clustering owing to inaccuracies in frame warping, thereby generating false positive errors. This issue is often caused by shifts in input data distribution and occlusion. However, our coherence-enhanced SFE, grounded in an online optimization solving detailed in Section 5.1, effectively counters this shift issue, helping to avoid low SFE accuracy and associated false positive errors. Secondly, despite adequate SFE, poor clustering accuracy can occur when real objects come too close to each other or when fake objects attach to real ones, resulting in merged object clusters and either false positive or negative errors. To mitigate this, we leverage extended historical frames to foster a broader understanding of the context, which aids in discerning the real objects from the fake ones more accurately. It is important to note that false negatives generally do not significantly impair navigation or trigger false alarms since a real object is indeed present, as discussed in the Failure Cases section. Lastly, when both SFE and clustering accuracy are satisfactory, ADoPTefficiently identifies attacks in the majority of cases, as evidenced by the data presented in Table 1.

On the other hand, while ADoPT operates independently of LiDAR object detection, immediate elimination of fake objects can enhance AV perception accuracy by reducing falsely detected instances.

# 7 Conclusion

We present the ADoPT framework, designed to detect LiDAR spoofing attacks on AVs by measuring temporal consistency at the point cloud level. ADoPT surpasses existing methods, delivering lower false positive rates and higher true positive rates. While currently focused on single-frame fake object injection attacks, ADoPT has the potential to address LiDAR spoofing attacks spanning consecutive frames by promptly eliminating detected spoofed points, converting them into benign frames, and continuously identifying spoofed objects in subsequent incoming frames. In the future, we aim to evolve ADoPT to counter diverse LiDAR data manipulation attacks (e.g., object removal attacks), thereby enhancing the robustness of perception modules in AVs.

# 8 Acknowledgements

We express our gratitude to Jiachen Sun and Qingzhao Zhang for their discussions on this work, and to our area chairs and anonymous reviewers for their invaluable feedback. This work is partially supported by NSF under grants CNS-1930041, CMMI-2038215, and Grant # 2112562 (the National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks).

# References

[1] Paul J. Besl and Neil D. McKay. A Method for Registration of 3-D Shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*, 1992.

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[3] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3D Tracking and Forecasting With Rich Maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative Perception for Connected Autonomous Vehicles Based on 3D Point Clouds. In *IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2019.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996.

[6] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A Point Set Generation Network for 3D Object Reconstruction From a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Zhongyuan Hau, Soteris Demetriou, Luis Muñoz-González, and Emil C. Lupu. Shadow-Catcher: Looking into Shadows to Detect Ghost Objects in Autonomous Vehicle 3D Sensing. In *Computer Security – ESORICS 2021 (ESORICS)*, 2021.

[8] Osamu Hirose. A Bayesian Formulation of Coherent Point Drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[9] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast Encoders for Object Detection From Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] Peng Li, Ruisheng Wang, Yanxia Wang, and Wuyong Tao. Evaluation of the ICP Algorithm in 3D Point Cloud Registration. *IEEE Access*, 2020.

[11] Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Neural Scene Flow Prior. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[12] YANG LI and Tatsuya Harada. Non-rigid Point Cloud Registration with Neural Deformation Pyramid. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[13] Jinshan Liu and Jung-Min Park. "Seeing is Not Always Believing": Detecting Perception Error Attacks Against Autonomous Vehicles. *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 2021.

[14] Xingyu Liu, Charles R. Qi, and Leonidas J. Guibas. FlowNet3D: Learning Scene Flow in 3D Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[16] Yanmao Man, Raymond Muller, Ming Li, Z. Berkay Celik, and Ryan Gerdes. That Person Moves Like A Car: Misclassification Attack Detection for Autonomous Systems Using Spatiotemporal Consistency. In *USENIX Security Symposium (USENIX Security)*, 2023.

[17] Himangi Mittal, Brian Okorn, and David Held. Just Go With the Flow: Self-Supervised Scene Flow Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] NVIDIA. Train With Mixed Precision, 2023. URL https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html.

[19] Jonathan Petit, Bas Stottelaar, Michael Feiri, and Frank Kargl. Remote Attacks on Automated Vehicles Sensors: Experiments on Camera and LiDAR. *Black Hat Europe*, 2015.

[20] Jhony Kaesemodel Pontes, James Hays, and Simon Lucey. Scene Flow from Point Clouds with or without Learning. In *International Conference on 3D Vision (3DV)*, 2020.

[21] Takami Sato, Yuki Hayakawa, Ryo Suzuki, Yohsuke Shiiki, Kentaro Yoshioka, and Qi Alfred Chen. Poster: Towards Large-Scale Measurement Study on LiDAR Spoofing Attacks against Object Detection. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022.

[22] Hocheol Shin, Dohyun Kim, Yujin Kwon, and Yongdae Kim. Illusion and Dazzle: Adversarial Optical Channel Exploits Against Lidars for Automotive Applications. In *Cryptographic Hardware and Embedded Systems – CHES (CHES)*, 2017.

[23] Jiachen Sun, Yulong Cao, Qi Alfred Chen, and Z. Morley Mao. Towards Robust LiDAR-based Perception in Autonomous Driving: General Black-box Adversarial Sensor Attack and Countermeasures. In *USENIX Security Symposium (USENIX Security)*, 2020.

[24] Jiachen Sun, Karl Koenig, Yulong Cao, Qi Alfred Chen, and Z Morley Mao. On Adversarial Robustness of 3D Point Cloud Classification under Adaptive Attacks. In *British Machine Vision Conference (BMVC)*, 2020.

[25] Jiachen Sun, Jiongxiao Wang, Weili Nie, Zhiding Yu, Zhuoqing Mao, and Chaowei Xiao. A Critical Revisit of Adversarial Robustness in 3D Point Cloud Recognition with Diffusion-Driven Purification. In *Proceedings of The International Conference on Machine Learning (ICML)*, 2023.

[26] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[27] Chaoyang Wang, Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Neural Prior for Trajectory Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[28] Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Density-aware Chamfer Distance as a Comprehensive Metric for Point Cloud Completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[29] Wenxuan Wu, Zhi Yuan Wang, Zhuwen Li, Wei Liu, and Li Fuxin. PointPWC-Net: Cost Volume on Point Clouds for (Self-)Supervised Scene Flow Estimation. In *Computer Vision – ECCV (ECCV)*, 2020.

[30] Chaowei Xiao, Ruizhi Deng, Bo Li, Taesung Lee, Benjamin Edwards, Jinfeng Yi, Dawn Song, Mingyan Liu, and Ian Molloy. AdvIT: Adversarial Frames Identifier Based on Temporal Consistency in Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[31] Qifan Xiao, Xudong Pan, Yifan Lu, Mi Zhang, Jiarun Dai, and Min Yang. Exorcising "Wraith": Protecting LiDAR-based Object Detector in Automated Driving System from Appearing Attacks. In *USENIX Security Symposium (USENIX Security)*, 2023.

[32] Chen Yan, Wenyuan Xu, and Jianhao Liu. Can You Trust Autonomous Vehicles: Contactless Attacks Against Sensors of Self-driving Vehicle. *Def Con*, 2016.

[33] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely Embedded Convolutional Detection. *Sensors*, 2018.

[34] Kaichen Yang, Tzungyu Tsai, Honggang Yu, Max Panoff, Tsung-Yi Ho, and Yier Jin. Robust Roadside Physical Adversarial Attack Against Deep Learning in Lidar Perception Modules. In *Proceedings of the ACM Asia Conference on Computer and Communications Security (ASIA CCS)*, 2021.

[35] Chengzeng You, Zhongyuan Hau, and Soteris Demetriou. Temporal Consistency Checks to Detect LiDAR Spoofing Attacks on Autonomous Vehicle Perception. In *Proceedings of the Workshop on Security and Privacy for Mobile AI (MAISP)*, 2021.

[36] Xumiao Zhang, Anlan Zhang, Jiachen Sun, Xiao Zhu, Y. Ethan Guo, Feng Qian, and Z. Morley Mao. EMP: Edge-Assisted Multi-Vehicle Perception. In *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2021.

[37] Yi Zhang, Zhiyu Xiang, Chengyu Qiao, and Shuya Chen. Accurate and Real-Time Object Detection Based on Bird's Eye View on 3D Point Clouds. In *International Conference on 3D Vision (3DV)*, 2019.