On the Adversarial Robustness of Temporal LiDAR Object Detectors

Mellon M. Zhang* Siddhant Panse* Akshal Dhal
Rishit Sarkar Zimo Fan Glen Chou
Georgia Institute of Technology
{meilongz,spanse30,adhal8,rsarkar44,zfan321,chou}@gatech.edu

Abstract

*Accurate 3D object detection is critical for autonomous driving, yet robustness to adversarial attacks – such as LiDAR spoofing – remains underexplored. While recent LiDAR-based detectors leverage temporal context to improve performance, they are only evaluated under benign conditions. In this work, we introduce **ATLAS** (Adversarial Temporal LiDAR Attack Suite), a large-scale, physically grounded benchmark for evaluating spoofing attacks on temporal detectors. ATLAS simulates realistic sensor-level attacks by injecting structured adversarial point clouds into multi-frame sequences. Our results reveal a counterintuitive trend: temporal models are consistently more vulnerable than single-frame baselines, with attack success rates increasing with memory length. We further observe systematic overconfidence on spoofed objects, indicating a failure to distinguish corrupted history from reliable inputs. Together, these findings show that temporal aggregation – when built on the assumption of clean observations – becomes a liability under adversarial perturbations, motivating the need for uncertainty-aware temporal perception.*

1. Introduction

Accurate 3D object detection is fundamental to reliable autonomous driving, requiring high precision and real-time performance in complex, dynamic environments. LiDAR sensors have become a cornerstone of modern perception systems due to their robustness to lighting variation and their ability to provide precise geometric and depth information beyond what RGB imagery alone can offer. As autonomous vehicles transition from controlled testing to large-scale real-world deployment, however, it becomes increasingly important to consider not only performance under benign conditions, but also robustness to anomalous and adversarial inputs. Among these, *LiDAR spoofing attacks* [11, 22] are particularly concerning: an external adversary can inject carefully crafted laser signals into the sensor, creating phantom objects without access to model internals, making such attacks feasible and increasingly realizable in real-world settings.

Despite this emerging threat, the development of LiDAR-based object detectors has largely been driven by performance on clean, curated datasets. Early approaches focused on single-frame detection [18, 30, 33, 34], processing each point cloud independently. More recently, temporal methods [2, 13, 36] have extended this paradigm by incorporating multi-frame context, aggregating information across time through feature fusion, proposal tracking, or memory mechanisms. These approaches consistently achieve state-of-the-art performance on standard benchmarks, demonstrating the benefits of temporal reasoning in improving detection accuracy and stability. However, both single-frame and multi-frame methods are predominantly evaluated under benign conditions, where observations are assumed to be accurate and free of adversarial corruption.

In this work, we argue that this assumption is fundamentally limiting. Existing temporal detectors are designed to exploit temporal consistency, but lack the ability to verify it. They treat both current observations and historical predictions as equally reliable, reusing memory without modeling uncertainty or detecting anomalies. While this design is effective under clean settings, it becomes fragile when the input stream is corrupted: errors can persist in memory, propagate across timesteps, and even be amplified through temporal aggregation. To systematically study this phenomenon, we introduce **ATLAS** (Adversarial Temporal LiDAR Attack Suite), the first large-scale, physically grounded benchmark for adversarial LiDAR spoofing built on the Waymo Open Dataset [26]. ATLAS simulates realistic sensor-level attacks by injecting structured adversarial point clouds into multi-frame sequences, enabling controlled evaluation of temporal detectors under distribution shift. Through extensive experiments, we show that temporal aggregation is not inherently more robust – in fact, models with longer memory are consistently more susceptible to spoofing, often underperforming single-frame baselines. We further find that these models exhibit systematic overconfidence on spoofed objects and fail to distinguish cor-

*Equal contribution.

rupted history from reliable signals. Together, our results reveal a fundamental limitation of existing multi-frame perception systems: memory, when unverified, acts as an amplifier of errors rather than a safeguard.

Our contributions are summarized as follows:

1. We introduce **ATLAS**, the first large-scale, physically grounded benchmark for adversarial LiDAR spoofing, enabling systematic evaluation of temporal object detectors under realistic multi-frame attacks.
2. We conduct a comprehensive empirical study of state-of-the-art temporal LiDAR detectors, showing that increased temporal context consistently leads to higher vulnerability and can degrade performance below single-frame baselines.
3. We identify key failure modes – overconfidence on adversarial inputs and error amplification through memory – and highlight the need for uncertainty-aware and reliability-driven temporal fusion mechanisms.

2. Related Work

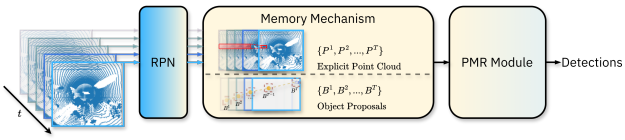

Figure 1. **State-of-the-art temporal LiDAR detectors follow a two-stage pipeline.** Proposals are first generated for the current frame via a region proposal network (RPN). These proposals query a memory module to retrieve temporal context, which is then used by a proposal merging and refinement (PMR) module to produce final detections for downstream use.

LiDAR Object Detectors. Most LiDAR object detectors operate in the single-frame setting, processing each point cloud independently across timesteps in a single forward pass [8, 19, 24, 30, 31, 33, 34, 35, 37, 38]. While effective, this design ignores the rich temporal structure present in driving scenarios. To address this limitation, multi-frame methods incorporate temporal context by aggregating information across time. Early temporal approaches adopt a single-stage design, aligning and fusing point clouds from consecutive frames to construct denser input representations [14, 31, 32]. However, these methods require accumulating multiple frames before inference, introducing substantial latency that limits their applicability in real-time systems. To mitigate this issue, recent work has shifted toward two-stage temporal architectures [2, 7, 9, 12, 13], highlighted in Figure 1. In these methods, a lightweight single-stage detector is used as a region proposal network (RPN) to first generate proposals from the current frame, which are then refined using features or predictions retrieved from past frames. This streaming formulation [36] enables immediate per-frame processing while maintaining a memory bank of historical context, striking a practical balance between latency and performance. Despite their strong performance on standard benchmarks [3, 26], the robustness of these temporal detectors remains poorly understood. In particular, their reliance on memory introduces new failure modes: errors in past predictions may propagate over time, and adversarial perturbations may accumulate or interact with stored history in nontrivial ways. How these memory mechanisms behave under adversarial conditions remains largely unexplored.

Physical LiDAR Spoofing. A growing body of work has established that LiDAR sensors are vulnerable to physical signal injection attacks that can fabricate phantom objects or suppress legitimate returns [15, 21, 25]. Early work by Sun et al. [25] demonstrated the feasibility of such attacks, showing that a custom spoofing device could inject up to 140 geometrically structured points and successfully mislead single-frame object detectors. Subsequent efforts have substantially advanced the capability, scale, and practicality of these attacks. PLA-LiDAR [15] scales injection to up to 4200 points, successfully inducing false detections in 51–98% of cases on widely used models such as PointPillars [18] and SECOND [30]. More recently, PhantomLidar [16] further increases realism and flexibility by injecting up to 16,000 points via electromagnetic interference, removing the need for precise knowledge of the target sensor's scanning pattern. Sato et al. [23] demonstrate the real-world implications of these attacks, showing that as few as 1,000 points arranged as a planar wall can trigger emergency braking in a full autonomous driving stack.

Notably, these works have kept pace with advances on the hardware attack surface, incorporating increasingly sophisticated injection mechanisms and relaxing attacker assumptions. However, their evaluation protocols remain largely anchored to early-generation, single-frame LiDAR object detectors [1, 5, 17, 18, 29, 30, 31]. As a result, the robustness of modern detection architectures, particularly recent single-stage improvements and temporal/multi-frame methods, has yet to be systematically evaluated. Furthermore, there is currently no large-scale, open-source benchmark that unifies these physically-grounded attack models and enables systematic evaluation across diverse detectors. This lack of a standardized evaluation infrastructure presents a critical barrier to developing and benchmarking robust LiDAR perception systems.

3. Adversarial Temporal LiDAR Attack Suite (ATLAS)

3.1. Threat Model

We consider a physically-realizable threat model for adversarial LiDAR perception shown in Figure 2: the roadside spoofer setting [23]. In this scenario, an attacker positioned on the roadside or a nearby vehicle emits synchronized laser signals to inject structured returns into the ego

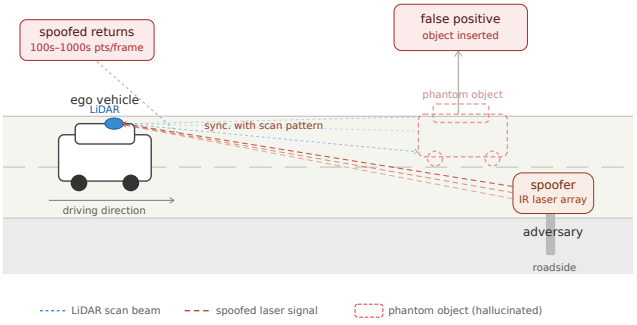

Figure 2. **Threat model.** A roadside adversary injects spoofed laser returns into the ego vehicle's LiDAR stream to induce phantom object detections.

LiDAR stream, creating false measurements. Prior work shows such systems can inject hundreds to thousands of points per frame [15, 16], sufficient to form coherent object-like geometries [25]. This threat model is both realistic and practical. It requires no access to model parameters, relying only on knowledge of LiDAR scan patterns, which is readily available for widely-used rolling-shutter sensors [3, 4, 10, 26]. Because the attack operates through direct physical interaction, it is difficult to detect or mitigate using standard cybersecurity defenses. Within this setting, we focus on object insertion attacks [25], where spoofed points induce the perception system to hallucinate non-existent objects. This is particularly dangerous as it directly alters scene understanding and can trigger unsafe behaviors such as unnecessary braking.

3.2. Benchmark Scope

We construct a large-scale spoofing benchmark that emulates the capabilities of modern physical LiDAR spoofing devices and systematically evaluates their impact on multi-frame detection pipelines. Our benchmark is built on the Waymo Open Dataset (WOD) [26], enabling evaluation under realistic driving scenarios with diverse traffic patterns and sensor conditions. To capture the design space of practical spoofing attacks, our benchmark varies three key factors: (1) point budget, (2) attack timing, and (3) adversary position. Permutations of these factors result in six variants of the WOD validation split. Altogether, ATLAS consists of $>$230k LiDAR point clouds, comprising 195 diverse driving trajectories each around 20 seconds long.

Point budget. We vary the number of injected points to probe detector sensitivity to spoofing density, considering three regimes: *easy* (100–200 points), *medium* (300–500 points), and *hard* (500–1000 points). The *easy* regime is aligned with current state-of-the-art physical spoofing systems, which can reliably generate geometrically coherent object-like patterns at this scale [25]. The *medium* and *hard* regimes are extrapolated stress-test settings designed to reflect plausible near-term attack capabilities. These regimes balance two competing factors: (i) maintaining geometric coherence of spoofed objects, and (ii) leveraging recent advances in high-density point injection, which can produce substantially larger point counts but are typically less structured [15, 16]. They also capture a realistic worst-case scenario in which multiple adversaries (e.g., 2–5 attackers) operate at the same time [28], effectively increasing the total injected point budget.

Attack timing. We model the temporal evolution of spoofing attacks by considering both their onset and accumulation. At attack onset, the detector operates on clean historical memory, allowing us to isolate the immediate impact of injected points. As the attack persists, spoofed observations are incorporated into the model's memory, enabling us to study how errors accumulate and propagate over time within multi-frame pipelines.

Adversary position. We consider two physically grounded attack modes that capture distinct adversary behaviors. In the *relative-position fixed* mode, the injected object maintains a constant pose relative to the ego vehicle, simulating an adversary that tracks or moves alongside the target vehicle [6]. In contrast, the *global-position fixed* mode places the spoofed object at a fixed location in the world frame, corresponding to a static roadside attacker that injects a phantom parked vehicle as the ego vehicle approaches it [21, 23, 25, 27].

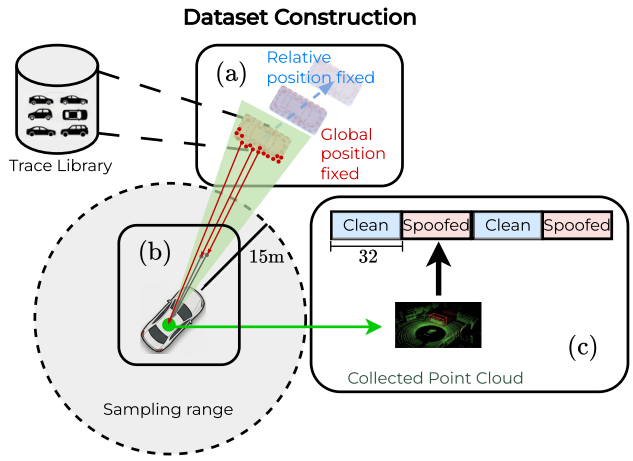

Figure 3. **Dataset Construction.** First, a centroid location is sampled in a range roughly 15 to 20m away from the ego vehicle. A vehicle trace collected from the Waymo Open dataset is randomly selected to be injected in that location. **(a)** Depending on the adversary position, the spoofed object either has its global position fixed (static adversary) or its position relative to the ego vehicle fixed (dynamic adversary). **(b)** The points are computed via spherical raytracing. **(c)** Lastly, the spoofing episodes are interspersed with clean episodes to measure the behavior of the perception models from attack onset to full memory corruption.

3.3. Benchmark Construction

Trace library. Rather than naively inserting synthetic points, we simulate attacks at the sensor level to better approximate physical signal injection [25]. We construct a library of real vehicle point cloud traces by extracting annotated vehicles from Waymo Open [26] and aggregating their LiDAR returns in a canonical vehicle coordinate frame. Each trace therefore consists of real sensor measurements from a physical object, preserving key properties such as range-dependent sparsity, occlusion patterns, and intensity distributions. By replaying these traces into new scenes, we can synthesize physically plausible phantom vehicles that closely resemble genuine LiDAR observations.

To generate an attack instance, we randomly sample a vehicle trace from this library and a target placement relative to the ego vehicle. Specifically, we sample a longitudinal offset $x \sim \mathcal{U}(15\text{m}, 20\text{m})$ and a lateral offset $y \sim \mathcal{U}(-3\text{m}, 3\text{m})$ relative to the ego vehicle. The sampled trace is first centered on its annotated bounding box, then perturbed with a small random yaw $\psi \sim \mathcal{U}(-0.2, 0.2)$ to introduce geometric variation. We then translate the trace to the sampled location, ground-align it, and assign range-consistent intensity and elongation statistics.

Modeling adversary position (Fig. 3a). In the *relative-position fixed* setting, the ghost vehicle is initialized at a sampled ego-relative offset (x_0, y_0) (see Trace Library) and maintains this offset throughout the spoof window. As a result, the injected object remains at a constant bearing and distance relative to the ego vehicle, mimicking an adversary that tracks or moves alongside it.

In the *global-position fixed* setting, the spoofed object is anchored in the world frame and does not move as the ego vehicle approaches it. We define an *anchor frame* as the final frame of each 32-frame spoof block. The object is first placed at the sampled ego-relative offset in this anchor frame and then transformed into global coordinates:

$$\mathbf{p}^{\text{world}} = \mathbf{R}_{\text{anc}}\mathbf{p}^{\text{ego}}_{\text{anc}} + \mathbf{t}_{\text{anc}}, \tag{1}$$

where $\mathbf{R}_{\text{anc}}$ and $\mathbf{t}_{\text{anc}}$ denote the ego-to-world rotation and translation. For each frame k within the spoof window, these world-frame points are projected back into the ego coordinate system:

$$\mathbf{p}^{\text{ego}}_k = \mathbf{T}_k^{-1}\mathbf{p}^{\text{world}}. \tag{2}$$

where $\mathbf{T}_k \in SE(3)$ is the ego-to-global transformation at frame k. As the ego vehicle moves, the ghost object therefore follows a natural approach-and-recede trajectory in the ego frame while remaining stationary in the world.

Sensor-level spoof simulation via raycasting (Fig. 3b). To enforce physical realism, we explicitly model occlusion through spherical raycasting. Both scene and spoofed points are projected into discretized spherical coordinates indexed by azimuth and beam. Within each bin, we resolve visibility via range ordering: spoof points behind all scene returns are discarded; spoof points in front occlude scene returns; and spoof points between existing returns remove only those farther along the ray. To approximate dual-return LiDAR behavior [20], we retain a secondary return with a small probability.

The resulting visible spoof and scene points are merged to form the final point cloud for each frame. Consequently, each frame contains either a clean observation or a physically consistent spoofed scene. When a spoofed object is present, its ground-truth bounding box is transformed into the current ego frame and recorded as the spoofed bounding box $g \in \mathbb{R}^8$ (comprising $[x, y, z, l, w, h, \theta, c]$ where θ is heading and c is the bounding box class). The spoofed bounding boxes are recorded for the downstream computation of attack success rate, defined in Sec. 4.2. All injected points are processed identically to real measurements, entering the pipeline prior to voxelization and undergoing the same downstream preprocessing.

Alternating spoof window (Fig. 3c). Temporal attack scheduling. To model sustained yet realistic attack behavior, we schedule spoofing in alternating temporal windows within each driving segment of the WOD validation split. Specifically, we partition each sequence into contiguous 32-frame blocks and alternate between clean and spoofed intervals. Let $s_i \in \{0, 1\}$ denote whether frame i is spoofed. We initialize with a clean frame ($s_0 = 0$), and for $i \geq 1$ define:

$$s_i = \mathbf{1}[\lfloor i/32 \rfloor \bmod 2 = 1]\,. \tag{3}$$

This produces a repeating pattern in which a phantom vehicle appears, persists for 32 frames (approximately 3.2 seconds), disappears, and then reappears after an equally long clean interval. We choose a 32-frame horizon for two complementary reasons. First, it aligns with the practical capabilities of existing LiDAR spoofing systems [27], which can sustain coherent signal injection over multi-second durations. Second, it matches the upper range of temporal context used by modern multi-frame detectors, whose memory lengths typically span 4–32 frames. By operating at this shared timescale, our protocol directly probes how adversarial signals interact with – and accumulate within – the full memory capacity of these models. This design enables us to study both the immediate impact of spoofing and its temporal propagation through multi-frame reasoning.

4. Evaluation Setup

We discuss our evaluated baselines in Sec. 4.1 and evaluation protocol for each baseline in Sec. 4.2.

4.1. Baselines

CenterPoint. CenterPoint [31] encodes a LiDAR point cloud into a bird's-eye-view (BEV) feature map using a

3D backbone, then applies convolution-based heads to produce a class-specific heatmap whose local peaks identify object centers. From each center, the network regresses 3D size, orientation, and velocity, unifying detection and tracking in a single anchor-free framework. A second-stage refinement extracts bilinear-interpolated face-center features from predicted bounding boxes to produce IoU-aware confidence scores and refined box estimates. CenterPoint 4-frame (hereafter referred to as CP4F) extends CenterPoint by concatenating four consecutive LiDAR sweeps into a single point cloud before encoding. This multi-frame aggregation produces denser voxel evidence, yielding richer representations particularly for stationary objects that accumulate points across frames.

MSF. MSF [12] is a two-stage temporal detector that uses CenterPoint 4-frame as its RPN. Current-frame proposals are propagated backward using estimated velocities, and 128 raw LiDAR points from each historical frame are pooled within expanding cylindrical regions centered on the propagated proposal locations. The sampled points are encoded via geometric and motion embeddings, then refined through three learning blocks containing multi-head self-attention and a Bidirectional Feature Aggregation (BiFA) module. BiFA enables cross-frame information exchange in both temporal directions, allowing persistent patterns across frames to be mutually reinforced. A query-based transformer decoder produces the final proposal representation for box regression and confidence prediction. In this work, we evaluate the performance of MSF with a four-frame point cloud history.

PTT. PTT [13] is a two-stage temporal detector that uses CP4F as its region proposal network. It requires only the current frame's point cloud alongside a 32 frame history of proposals, avoiding storage of historical point clouds. Point-to-Proposal features encode geometric relationships between current-frame points and historical proposal box corners, while Proposal-to-Proposal features capture inter-frame displacements. These are combined into per-frame point-trajectory features. These features are split into long-term and short-term memories, where short-term features query the long-term memory via cross-attention to extract relevant historical trends. A future-aware module further encodes anticipated trajectory features from estimated velocities. Finally, a point-trajectory aggregator fuses all memory outputs with current-frame point features to produce final detections.

4.2. Evaluation Protocol

Proposal generation with CenterPoint. To ensure that spoofed observations propagate through the detection pipeline realistically, we recompute region proposals using CP4F [31].

First, spoofed points are ray-casted into the four-frame history of a current frame based on the spoof schedule described by Eq. (3). Then, for each frame t, a four-frame point cloud is constructed by transforming the single-frame point clouds from frames $t-3, ..., t$ into the ego coordinate frame at time t. This can be represented by Eq. (4),

$$\mathcal{P}^{(t)} = \bigcup_{j=0}^{3} \left\{ \left[\mathbf{T}_t^{-1} \mathbf{T}_{t-j} \tilde{\mathbf{p}},\ p_{3:5},\ 0.1j \right] : \mathbf{p} \in \mathcal{P}_{t-j} \right\}, \quad (4)$$

where $\mathbf{T}_k \in SE(3)$ is the ego-to-global transformation at frame k, $\tilde{\mathbf{p}} = [p_x, p_y, p_z, 1]^\top$ is the homogeneous point coordinate, $p_{3:5}$ are the intensity and elongation channels, and $0.1j$ is a lag timestamp encoding the frame's temporal offset in seconds. When all four frames in the history are clean, the original unmodified dataloader points are used; for frames containing spoofed objects in the history, the modified aggregated cloud is constructed. A CenterPoint forward pass is performed, and the resulting proposals, labels, and confidence scores are stored per-frame for use as RPN input to the temporal detectors.

MSF Evaluation. For each frame t, the four-frame point cloud is reconstructed from the spoofed dataset following Eq. (4). The stored CenterPoint proposals undergo a velocity rescaling identical to the one used by the official MSF pipeline, as shown in Eq. (5):

$$(\hat{v}_x, \hat{v}_y) = (-0.1 \cdot v_x,\ -0.1 \cdot v_y). \quad (5)$$

This converts the CenterPoint proposal velocities into backward motion used by MSF's motion-guided point sampler to propagate proposals to their estimated past locations. The transformed bounding box proposals, scores, labels, and four-frame point cloud are used for an MSF forward pass.

PTT evaluation. For each frame t, the stored proposals from the preceding 32 frames are collected, velocity-rescaled via Eq. (5), and transformed into the ego coordinate frame of frame t:

$$\hat{\mathbf{B}}_k = f_{\text{box}}\left(\mathbf{T}_t^{-1}\mathbf{T}_k,\ \mathbf{B}_k\right), \quad k = t-31, \ldots, t, \quad (6)$$

where f_{box} denotes the rigid-body transformation of box center, heading, and velocity. A PTT forward pass is then performed, using the current frame's point cloud and transformed proposal history, label history, and confidence score history as inputs.

Evaluation Metric. We report attack success rate (ASR) to quantify the effectiveness of spoofing attacks on each baseline model. Let g denote the ground-truth bounding box of the spoofed (phantom) object. A predicted bounding box is considered a *successful false positive* if it overlaps with g with intersection-over-union (IoU) greater than or equal to a threshold b. Following common practice for vehicle detection, we use $b = 0.7$ [26].

A spoofing attack is deemed successful if it induces at least one such false positive detection corresponding to the

Mode	CenterPoint	CenterPoint 4F	MSF	PTT
memory length (# frames)	1	4	4	32
rel. fixed easy	0.11	0.12	0.13	0.14
rel. fixed medium	0.12	0.22	0.29	0.26
rel. fixed hard	0.10	0.29	0.35	0.35
global easy	0.16	0.18	0.19	0.21
global medium	0.30	0.36	0.41	0.44
global hard	0.32	0.45	0.46	0.50

Table 1. Vehicle attack success rate (ASR) across detectors and attack modes.

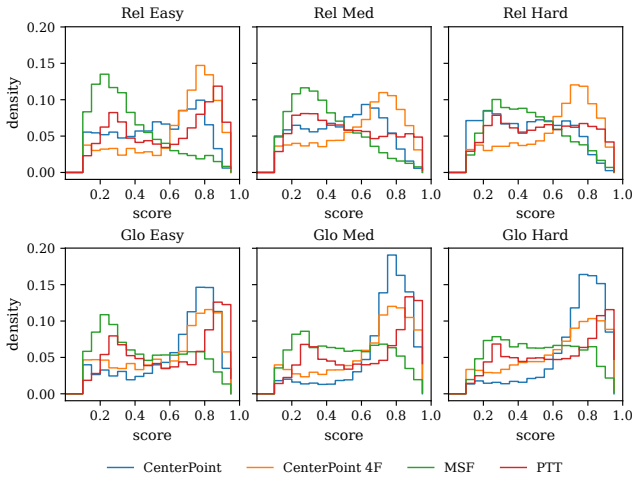

Figure 4. Spoof detection score distributions across detectors and attack modes.

spoofed target. The ASR is then defined as the fraction of spoofed targets for which a successful false positive is produced:

$$\text{ASR} = \frac{\#\text{ successful spoofed targets}}{\#\text{ total spoofed targets}}. \tag{7}$$

In addition to ASR, we analyze the confidence scores of the induced phantom detections to better understand model calibration under attack.

5. Results

We structure our results section to highlight four critical findings from our evaluations on ATLAS. For additional analysis and detailed discussion of these results, please see Sec. 6.

1. Temporal aggregation amplifies adversarial errors rather than correcting them (Finding 1).
2. Temporal detectors exhibit systematic overconfidence on adversarial inputs (Finding 2).
3. Temporal vulnerability saturates rather than growing monotonically with perturbation duration (Finding 3).
4. Temporally coherent adversarial signals are most effective (Finding 4).

Finding 1: Temporal aggregation amplifies adversarial errors rather than correcting them. Table 1 shows that

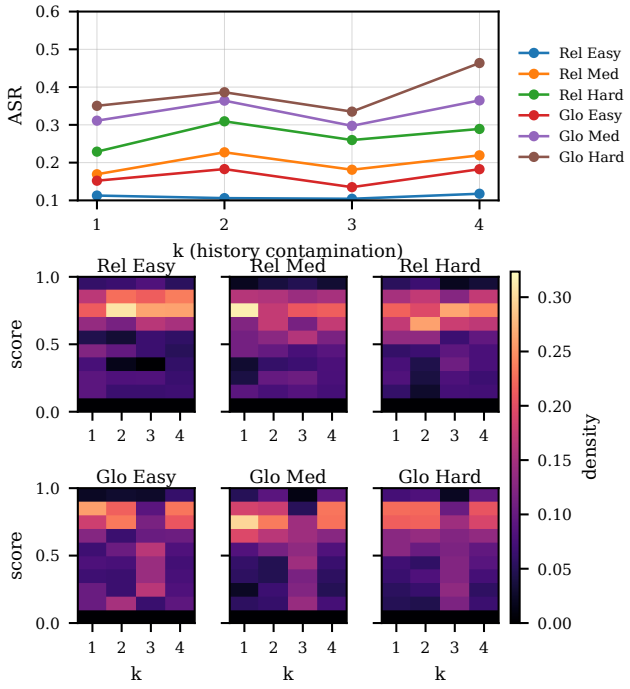

Figure 5. CP4F ASR (top) and confidence score distributions vs. number of spoofed historical frames in the 4-frame input window (bottom).

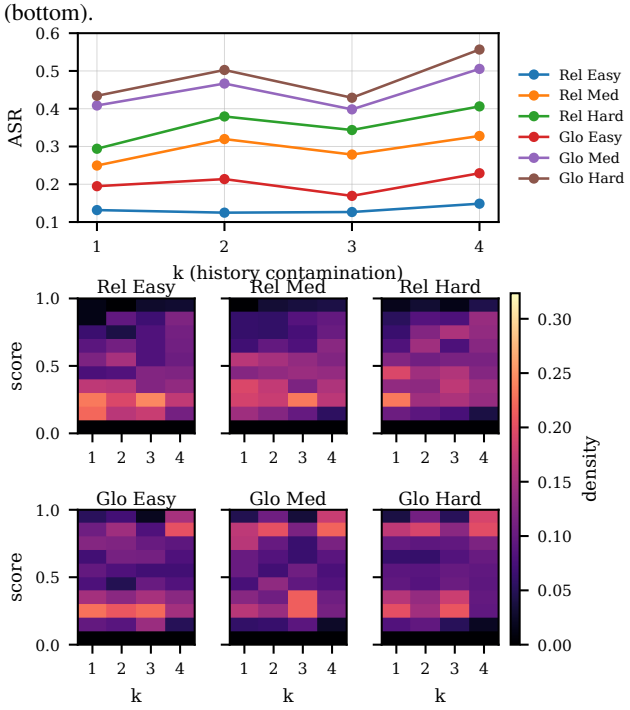

Figure 6. MSF ASR (top) and confidence score distributions vs. number of spoofed historical frames in the 4-frame input window (bottom).

robustness degrades as temporal context increases: while longer memory improves clean performance, it consistently increases vulnerability under spoofing. CenterPoint (single-frame) [31] is the most robust, followed by CP4F and MSF

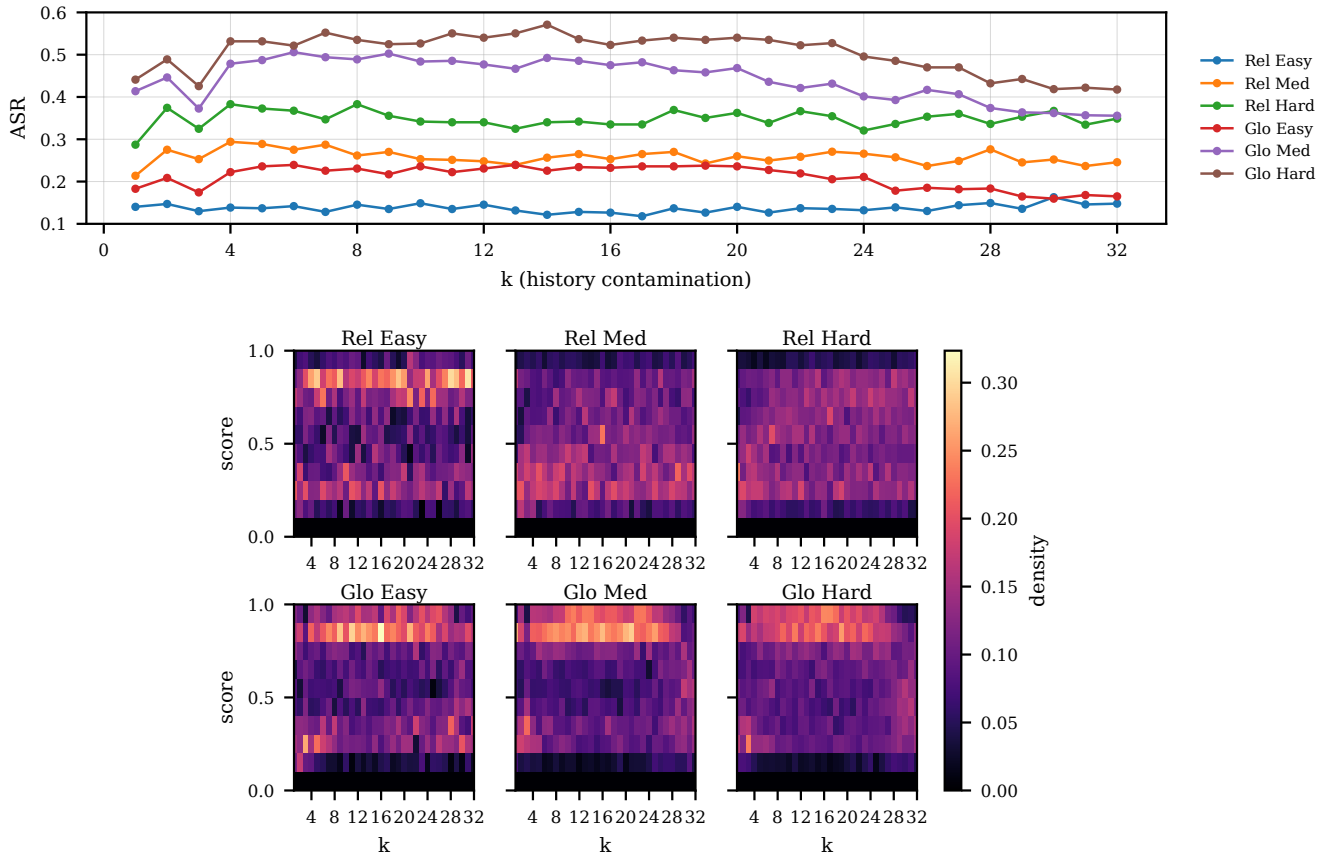

Figure 7. PTT history contamination analysis. **Top:** ASR vs. number of spoofed frames k in the 32-frame proposal trajectory. **Middle/Bottom:** Per-k score distributions of successful phantom detections under relative-fixed (middle) and global-fixed (bottom) placement.

[12] (4 frames), and PTT [13] (32 frames). Under the global-hard setting (see Sec. 3.2), ASR increases from 0.32 (CenterPoint) to 0.45 (CP4F), 0.46 (MSF), and 0.50 (PTT).

This inversion suggests that temporal refinement does not act as a corrective signal under distribution shift. Instead, memory mechanisms reinforce adversarial structure: once a spoofed object is introduced, multi-frame models accumulate and stabilize it over time, producing coherent and well-localized false positives.

Finding 2: Temporal detectors exhibit systematic overconfidence on adversarial inputs. Figure 4 shows that most detectors assign high confidence to spoofed objects, often comparable to real detections. In the relative-fixed setting, CP4F and CenterPoint exhibit pronounced peaks near $\sim$0.8, while PTT assigns high confidence to roughly 50% of spoofed predictions. Under global-fixed placement, CP4F, CenterPoint, and PTT again show strong high-confidence modes. MSF is the only exception, producing mostly low-confidence predictions, though even it assigns up to $\sim$20% of detections to the 0.7–0.8 range in global settings. This reveals a critical failure of uncertainty calibration: temporal pipelines implicitly treat both current observations and stored history as trustworthy. As a result, spoofed proposals are not only preserved but also reinforced with high confidence, leaving downstream modules with no signal to reject them. Second-stage refinement [12, 13] reshapes score distributions but does not reliably correct this miscalibration. We discuss possible explanations in Sec. 6.

Finding 3: Temporal vulnerability saturates rather than growing monotonically with perturbation duration. Figures 5, 6, and 7 show that ASR varies non-monotonically with the number of perturbed historical frames k. While temporal detectors with longer memory are more vulnerable overall (Finding 1), increasing the duration of perturbation within a fixed model does not consistently strengthen attacks. CP4F and MSF exhibit only marginal gains (at most $\sim$10% under global-hard), while PTT increases up to $k = 16$ before declining as k grows further. The score heatmaps reveal a consistent pattern: intermediate values of k capture most high-confidence false positives, while larger k yields diminishing or reduced confidence (e.g., the high-confidence band in PTT at $k \sim 5$–25 dissipates at larger k). This indicates that temporal detectors do not accumulate evidence indefinitely. Instead, spoofed signals are reinforced only up to a point, after which repeated observations provide little new information and can dilute the representation used for refinement.

Taken together, these results clarify that the vulnerability

introduced by temporal memory (Finding 1) is fundamentally bounded: longer memory increases a model's ability to reinforce adversarial signals, but within a fixed model this reinforcement saturates, so additional corrupted history does not yield unbounded gains in attack success.

Finding 4: Temporally coherent adversarial signals are most effective. Globally fixed spoofing consistently yields higher ASR than relative-fixed placement across all detectors and budgets (Table 1). Under the hard setting, global placement achieves ASRs of 0.32, 0.45, 0.46, and 0.50 for CenterPoint, CP4F, MSF, and PTT, compared to 0.10, 0.29, 0.35, and 0.35 under relative placement, with the largest gap observed for CP4F (+0.16). This highlights a key vulnerability of temporal detectors: they are optimized to exploit cross-frame consistency. Globally fixed spoofs mimic this consistency and are therefore reinforced over time, whereas relative placement introduces motion inconsistency that partially breaks temporal alignment. Upping the point budget further boosts ASR across nearly all settings, but does not systematically increase confidence (Fig. 4). This decoupling suggests that attack success is driven by temporal coherence rather than raw signal strength.

6. Discussion

We discuss the key findings from Sec. 5, trace them to baseline architectural details, and discuss potential avenues for future work motivated by these observations.

ASR as a function of memory length and perturbation duration. As observed in (Finding 1) and (Finding 3), we uncover a consistent duality in how temporal context impacts robustness. Across models, longer memory increases susceptibility to spoofing, as it enables adversarial signals to persist and be reinforced over time. However, within a fixed model, this effect is not unbounded: attack success saturates with perturbation duration, and ASR does not increase monotonically as more of the memory is corrupted. This suggests that temporal aggregation benefits spoofing only insofar as it introduces new supporting evidence. Once a spoofed object has been sufficiently integrated into memory, additional repeated observations provide diminishing returns and can even weaken the attack. While temporal coherence remains important – explaining why globally fixed spoofs are more effective than relative ones – even perfectly coherent and physically plausible signals (e.g., a static parked car) exhibit this saturation effect. One possible explanation is that temporal detectors are implicitly biased toward informative or evolving signals. In practice, training data is dominated by dynamic scenes, which may lead models to down-weight objects that remain unchanged over long horizons. Consistent with this, we observe that static spoofs peak in effectiveness around $k \approx 16$, after which ASR declines, suggesting that overly persistent signals fall outside the model's effective operating regime.

Higher ASR on multi-frame detectors vs CenterPoint. As mentioned in Sec. 4.1, CenterPoint 4-frame [31] feeds an aggregated 4-frame point cloud into the standard CenterPoint pipeline. When assumptions of clean frames become violated, this means that spoofed points can receive up to $4\times$ densification, which becomes especially strong for the global-fixed case. This produces stronger voxel features and more pronounced heatmap peaks than any single frame alone, directly increasing ASR. Since CP4F serves as the first stage of MSF and PTT, this amplified phantom proposal propagates downstream.

Lower confidence scores from MSF. MSF [12] propagates CP4F's phantom proposals backward in time and samples 128 raw LiDAR points per proposal per frame within expanding cylindrical regions. The BiFA module reinforces the resulting features bidirectionally, increasing the likelihood that the phantom objects survive to final detection. However, the expanding cylinders also capture ground and background points alongside phantom returns, potentially diluting the signal. This may explain why MSF's confidence scores remain low-to-moderate (0.2-0.4) despite its higher ASR: its region-based network, trained on the coherent point patterns of real objects, may map these mixed quality features to lower confidence even when temporal consistency is high.

Designing future methods. Our findings highlight a critical weakness in existing temporal object detectors. These models are designed to exploit temporal consistency but lack the ability to verify it. As a result, memory acts as an amplifier of errors rather than a safeguard, allowing adversarial signals to persist over time. Notably, none of the evaluated detectors incorporate mechanisms to assess the trustworthiness of their temporal inputs, treating all historical evidence as equally reliable. These findings motivate several directions for future work. Uncertainty-aware temporal fusion could explicitly down-weight low-confidence or out-of-distribution history, preventing unreliable signals from dominating the refinement process. Alternatively, anomaly detection modules could identify and suppress temporally inconsistent proposals before they propagate through memory. Overall, our results suggest a need to rethink temporal modeling not as unconditional aggregation, but as selective and reliability-aware reasoning over time.

7. Conclusion

We introduce ATLAS, a physically grounded LiDAR spoofing benchmark, and evaluate temporal 3D object detectors under adversarial perturbations. Our findings suggest that current temporal detectors assume reliable history and fail when this assumption is violated, highlighting the need for next-generation approaches that incorporate uncertainty-aware temporal reasoning.

Acknowledgements. This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.

References

[1] Temesgen Mikael Abraha, John Brandon Graham-Knight, Patricia Lasserre, Homayoun Najjaran, and Yves Lucet. Convnet-generated adversarial perturbations for evaluating 3d object detection robustness. *Sensors*, 25(19):6026, 2025. 2

[2] Ben Agro, Sergio Casas, Patrick Wang, Thomas Gilles, and Raquel Urtasun. Mad: Memory-augmented detection of 3d objects. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1449–1460, 2025. 1, 2

[3] Wilson Benjamin, Qi William, Agarwal Tanmay, Lambert John, Singh Jagjeet, Khandelwal Siddhesh, Pan Bowen, Kumar Ratnesh, Hartnett Andrew, Kaesemodel-Pontes Jhony, Ramanan Deva, Carr Peter, and Hays James. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2, 3

[4] 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 *CVPR*, 2020. 3

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

[6] Yulong Cao, S Hrushikesh Bhupathiraju, Pirouz Naghavi, Takeshi Sugawara, Z Morley Mao, and Sara Rampazzi. You can't see me: Physical removal attacks on lidar-based autonomous vehicles driving frameworks. In *32nd USENIX security symposium (USENIX Security 23)*, pages 2993–3010, 2023. 3

[7] Xuesong Chen, Shaoshuai Shi, Benjin Zhu, Ka Chun Cheung, Hang Xu, and Hongsheng Li. Mppnet: Multi-frame feature intertwining with proxy points for 3d temporal object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[8] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *CVPR*, 2023. 2

[9] Chenxu Dang, Zaipeng Duan, Pei An, Xinmin Zhang, Xuzhong Hu, and Jie Ma. Faster: Focal token acquiring-and-scaling transformer for long-term 3d object detection, 2025. 2

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 3

[11] Amira Guesmi and Muhammad Shafique. Navigating threats: A survey of physical adversarial attacks on lidar perception systems in autonomous vehicles, 2024. 1

[12] Chenhang He, Ruihuang Li, Yabin Zhang, Shuai Li, and Lei Zhang. Msf: Motion-guided sequential fusion for efficient 3d object detection from point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5196–5205, 2023. 2, 5, 7, 8

[13] Kuan-Chih Huang, Weijie Lyu, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Ptt: Point-trajectory transformer for efficient temporal 3d object detection. In *CVPR*, 2024. 1, 2, 5, 7

[14] Rui Huang, Wanyue Zhang, Abhijit Kundu, Caroline Pantofaru, David A Ross, Thomas Funkhouser, and Alireza Fathi. An lstm approach to temporal 3d object detection in lidar point clouds. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[15] Zizhi Jin, Xiaoyu Ji, Yushi Cheng, Bo Yang, Chen Yan, and Wenyuan Xu. Pla-lidar: Physical laser attacks against lidar-based 3d object detection in autonomous vehicle. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1822–1839. IEEE, 2023. 2, 3

[16] Zizhi Jin, Qinhong Jiang, Xuancun Lu, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. Phantomlidar: Cross-modality signal injection attacks against lidar. In *NDSS*, San Diego, CA, USA, 2025. 2, 3

[17] Ryoma Kobayashi, Kazuki Nomoto, Yuta Tanaka, Goki Tsuruoka, and Tatsuya Mori. Invisible but detected: Physical adversarial shadow attack and defense on LiDAR object detection. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 7369–7386. USENIX Association, 2025. 2

[18] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 1, 2

[19] Jinyu Li, Chenxu Luo, and Xiaodong Yang. Pillarnext: Rethinking network designs for 3d object detection in lidar point clouds. In *CVPR*, 2023. 2

[20] Y. Li et al. Detection and utilization of reflections in lidar scans for autonomous systems. *Sensors*, 2024. 4

[21] Takami Sato, Yuki Hayakawa, Ryo Suzuki, Yohsuke Shiiki, Kentaro Yoshioka, and Qi Alfred Chen. Practical removal attacks on lidar-based object detection in autonomous driving. In *Symposium on Vehicle Security and Privacy*, San Diego, CA, USA, 2023. 2, 3

[22] Takami Sato, Yuki Hayakawa, Ryo Suzuki, Yohsuke Shiiki, Kentaro Yoshioka, and Qi Alfred Chen. Lidar spoofing meets the new-gen: Capability improvements, broken assumptions, and new attack strategies. In *Proceedings 2024 Network and Distributed System Security Symposium*. Internet Society, 2024. 1

[23] Takami Sato, Ryo Suzuki, Yuki Hayakawa, Kazuma Ikeda, Ozora Sako, Rokuto Nagata, Ryo Yoshida, Qi Alfred Chen, and Kentaro Yoshioka. On the realism of lidar spoofing attacks against autonomous driving vehicle at high speed and long distance. In *NDSS*, 2025. 2, 3

[24] Guangsheng Shi, Ruifeng Li, and Chao Ma. Pillarnet: Real-time and high-performance pillar-based 3d object detection. In *ECCV*, 2022. 2

[25] 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, 2020. 2, 3, 4

[26] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1, 2, 3, 4, 5

[27] Ryo Suzuki, Takami Sato, Yuki Hayakawa, Kazuma Ikeda, Ozora Sako, Rokuto Nagata, Qi Alfred Chen, and Kentaro Yoshioka. Wip: Towards practical lidar spoofing attack against vehicles driving at cruising speeds. In *ISOC Symposium on Vehicle Security and Privacy (VehicleSec)*, 2024. 3, 4

[28] Masoud Jamshidiyan Tehrani, Marco Gabriel, Jinhan Kim, and Paolo Tonella. Dynamic deception: When pedestrians team up to fool autonomous cars, 2026. 3

[29] James Tu, Mengye Ren, Siva Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection, 2020. 2

[30] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. In *Sensors*, 2018. 1, 2

[31] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, 2021. 2, 4, 5, 6, 8

[32] Rui Yu, Runkai Zhao, Cong Nie, Heng Wang, HuaiCheng Yan, and Meng Wang. Listm: Boosting 3d object detection with temporal motion estimation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2024. 2

[33] Gang Zhang, Junnan Chen, Guohuan Gao, Jianmin Li Li, and Xiaolin Hu. HEDNet: A hierarchical encoder-decoder network for 3d object detection in point clouds. In *NeurIPS*, 2023. 1, 2

[34] Gang Zhang, Junnan Chen, Guohuan Gao, Jianmin Li, Si Liu, and Xiaolin Hu. SAFDNet: A simple and effective network for fully sparse 3d object detection. In *CVPR*, 2024. 1, 2

[35] Guowen Zhang, Lue Fan, Chenhang He, Zhen Lei, Zhaoxiang Zhang, and Lei Zhang. Voxel mamba: Group-free state space models for point cloud based 3d object detection. *arXiv preprint arXiv:2406.10700*, 2024. 2

[36] Mellon M. Zhang, Glen Chou, and Saibal Mukhopadhyay. Towards streaming lidar object detection with point clouds as egocentric sequences, 2025. 1, 2

[37] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, 2018. 2

[38] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *ECCV*, 2022. 2