J Intel Connect Veh

Research Article

Transformer-based out-of-distribution (OOD) input detection for trustworthy self-parking

Hadi Ballout[1], Riccardo Berta[1, ✉], Luca Lazzaroni[1], Alessandro Pighetti[1,2], Matteo Fresta[1], Ammar Saad[1], Vahid Hashemi[3], Akshay Dhonthi[3], Francesco Bellotti[1]

Cite this article: https://doi.org/10.26599/JICV.2026.9210093

ABSTRACT: The deployment of automated driving (AD) systems demands trustworthiness, explainability, and accountability. A challenge is detecting out-of-distribution (OOD) inputs from sensor noise, hardware failures, adversarial interference, or deviations from the operational design domain. Common approaches use OOD samples during training or threshold selection, potentially limiting generalization beyond the anomaly types represented during development. We argue that OOD-unaware novelty detection is crucial for detection under unforeseen conditions. This paper enhances a deep-reinforcement-learning-based self-parking agent developed in CARLA by integrating a novel OOD detector. We present Time-Frequency-Memory-enhanced (TFMe), a dual-branch, memory-augmented, encoder-only Transformer that outperforms the evaluated baselines. In OOD-unaware settings, decision thresholds are calibrated exclusively on in-distribution validation data using a 99th-percentile rule. We assess sensitivity across multiple noise, attack types, and severity levels. Results show that OOD-unaware calibration maintains consistent performance across anomaly types and intensities, avoiding the drops observed under OOD-aware calibration. Experiments on real-world data from Lyft with synthetically injected anomalies further show that OOD-unaware calibration transfers more effectively to the evaluated unseen anomaly types. Transferability across parking environments is assessed on an unseen 60° angled layout without retraining. We show that the TFMe-enhanced ADF mitigates collision risk and reduces timeouts in the evaluated parking scenarios by issuing Take-Over Requests after repeated anomaly detections. TFMe has lower inference latency than the self-parking agent, allowing it to operate in parallel without extending the estimated critical path. Finally, compared with Raspberry Pi 5, deployment on a Jetson Orin Nano achieves 19% lower inference time and 30% lower energy consumption.

KEYWORDS: Automated driving, Automotive deployability, Deep Reinforcement Learning (DRL), Operational Design Domain (ODD), Out-of-Distribution (OOD) data, SHapley Additive exPlanations (SHAP), Take-over request (TOR), Transformer architecture.

1 Introduction

There is increasing recognition-both in industry and regulatory frameworks-that real-world deployment of automated driving (AD) systems requires not only high performance in conventional machine learning metrics but also demonstrably trustworthy, predictable, and explainable system-wide behavior (Gyllenhammar et al., 2025), (Hoppe et al., 2024). This has driven research into formal verification of neural architectures (Cheng et al., 2019), adversarial search (Henriksen and Lomuscio, 2020), and explainability methods (Bellotti et al., 2023), with particular emphasis on detecting out-of-distribution (OOD) inputs, that deviate from the distribution used for training (i.e., in-distribution (ID) data). OOD inputs arise from sensor noise (Y. Zhang et al., 2024), failures (Rodríguez-Arozamena et al., 2025), adversarial attacks (Girdhar et al., 2023), or operational design domain (ODD) breaches. Runtime monitoring of OOD data is critical to identify anomalies (Javed et al., 2021) that could compromise system integrity (Ayerdi et al., 2024) and provide traceability for system improvement and accountability in compliance with emerging safety and liability frameworks (Kacianka and Pretschner, 2021; Rowe et al., 2024).

This paper addresses the problem of robustness in OOD input detection by tackling two main critical gaps in the state-of-the-art. Most existing anomaly detection approaches rely on sophisticated supervised or unsupervised techniques and achieve impressive results (Xu et al., 2022). However, these methods typically incorporate some form of explicit knowledge about detectable anomalies during model development, at least through exposure to OOD samples for hyperparameter tuning or validation (van Wyk et al., 2020). While reasonable, given that many noise patterns and attack vectors are known, this practice inherently constrains detection capabilities to predefined anomaly types and intensity levels (Sun et al., 2022), limiting generalization to truly unforeseen conditions. This limitation motivates the investigation of a complementary paradigm based on OOD-unaware (pure) novelty detection, which requires no exposure to OOD data in training (included tuning).

Formally, the novelty detection problem is defined as follows. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space, and P_0 be an unknown probability distribution over $\mathcal{X}$ representing normal data. Given n independent and identically distributed (i.i.d.) samples distributed according to P_0, $\{x_i\}_{i=1}^n \sim P_0$ the objective is to learn a decision rule f such that for a new sample x,

[1] Department of Electrical, Electronic and Telecommunication Engineering (DITEN), University of Genoa, Genoa 16145, Italy. [2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK. [3] AUDI AG, Ingolstadt 85057, Germany.

✉ Corresponding author. E-mail: riccardo.berta@unige.it

Received: April 16, 2026; Revised: August 3, 2026; Accepted: August 7, 2026

(a) (b)

Fig. 1. Snapshot of a parking maneuver in CARLA (a) and bird's-eye view of a simulation episode (b).

$$f(x) = 1 \text{ iff } x \nsim P_0 \quad (1)$$

In the OOD-unaware novelty detection setting, no samples from any anomalous or alternative distributions are available during training nor validation. The underlying idea is to learn a model of normality from one-class data training and, in inference, detect any statistically significant deviation from the learned distribution.

Although novelty detection is a well-established machine learning task, its systematic application to decision-making in safety-critical autonomous systems is still relatively unexplored. This work addresses the lack of rigorous quantitative comparisons between knowledge-based (OOD-aware) and pure novelty (OOD-unaware) detection approaches (this is the methodology comparison gap), with the broader objective of advancing the robustness of autonomous driving systems under unforeseen conditions (safety-critical innovation gap).

We propose a novelty detector model featuring an encoder-only Transformer architecture integrated with a memory module to mitigate over-generalization, and a frequency domain component to capture multi-scale contextual dependencies for improved in-distribution (ID) data pattern representation.To the best of our knowledge, no previous Transformer architecture combines a dual (time/frequency) encoder with memory modules for both domains and a disjunctive (OR-gate) decision rule (cf. Section 3.2). We benchmark this model against two state-of-the-art frameworks: Adversarial Transformer-Based Anomaly Detection (ATUAD) (Yu et al., 2025) and Adversarial Time-Frequency Reconstruction Network (ATF-UAD) (Fan et al., 2023).

As another key innovation, we propose the integration of an OOD input detection module into an industry-oriented research automated driving function (ADF). The ADF features a state-of-the-art deep reinforcement learning (DRL) decision-making agent that enables end-to-end, map-less local path planning and trajectory tracking within a CARLA-simulated parking environment (Lazzaroni et al., 2024). The vehicular agent executes precise low-speed maneuvers in a challenging parking scenario featuring comb-shaped layouts with randomly distributed vehicles near the target parking lot (Fig. 1).

This research was conducted as part of the ********** industrial research project, a consortium comprising leading original equipment manufacturers (OEMs) and automotive suppliers. Hi-Drive seeks to advance AD functions (ADFs) through the development of innovative technological enablers across multiple domains, including connectivity and digital infrastructure, precise positioning, cybersecurity, and machine learning (Capello et al., 2023). Within this framework, the proposed OOD detection module represents a key enabling technology, specifically designed to enhance the reliability of the self-parking ADF.

As part of the work, we have developed LiDARacks, a Python toolkit to generate four types of physically-grounded LiDAR anomalies for robust sensor testing: Background noise (simulating adverse weather conditions (Yu et al., 2024)), Zero-Range Background (ZRB, modeling sensor obstruction), Electromagnetic Interference (EMI, representing electronic noise (Liu et al., 2024)), and Occlusion (ray diffusion/absorption). Inspired by the framework proposed in (Sun et al., 2022), which features 15 corruption types, we focused on the most sensor-relevant anomalies and improved physical fidelity by replacing Gaussian noise with EMI, implementing occlusion through blinded ray counts rather than FOV angles, and optimizing background noise generation to focus on observable regions (Wang et al., 2024) for accurate severity calibration. In LiDARacks, all anomaly levels are tied to concrete sensor metrics, ensuring industrial relevance while maintaining configurable testing conditions that preserve physical plausibility across various degradation scenarios.

Targeting industrial applications, we apply the OOD input detector to the self-parking decision-making agent developed in Hi-Drive, which generates values for four driving actions (steering angle, throttle, brake, and reverse) (Lazzaroni et al., 2024). Both modules process five consecutive frames of LiDAR observations and vehicle signals. Fig. 2 shows our enhanced ADF system architecture, in which the OOD detections are fed into the Take-Over Request (TOR) manager. The TOR is a critical feature of SAE Level 3/4 driving automation (SAE International, 2014), through which an ADF prompts the driver to intervene when encountering scenarios beyond its operational limits. While TORs have been extensively studied in contexts such as ODD limit violations (Lee et al., 2020) or system malfunctions (e.g., (Soares et al., 2021)), literature lacks deployable in-vehicle architectures that integrate anomaly detection with TOR, also beyond self-parking ADFs.

In a pure novelty detection perspective, we train the OOD detector exclusively on ID data and optimize its hyperparameters to minimize false positives, which could otherwise undermine user acceptance. To interpret the detector's behavior, we conduct a global SHAP analysis ((Lundberg and Lee, 2017; Shapley, 1952)), a widely adopted explainable AI technique in the industrial power sector (Trivedi et al., 2024). Finally, we evaluate the ADF's deployment in automotive settings by assessing embedded performance metrics across two resource-constrained platforms, the AI-oriented Jetson Orin Nano, and Raspberry Pi 5.

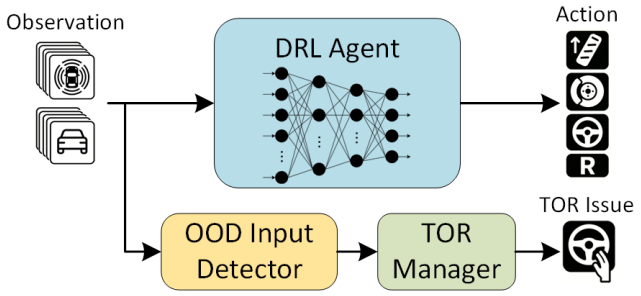

Fig. 2. High-level architecture of the enhanced ADF.

The remainder of the paper is organized as follows. Section 2 presents the related work, Section 3 introduces the methodology, while the evaluation of the proposed system is discussed in Section 4. Section 5 draws the conclusions on the work done and suggests directions for future research.

2 Related work

The development of an anomaly detection module encompasses two primary dimensions: anomaly modeling and implementation, followed by the design and implementation of the machine learning (ML)-based detection system. This section is structured accordingly.

2.1. Anomaly Modeling and Implementation

Anomaly modeling and implementation is the systematic process of characterizing deviations from a sensor's normal behavior. In the AD domain, recent studies have investigated the effects of LiDAR sensor inaccuracies on automated driving systems (ADS), revealing significant implications for perception and planning tasks. As demonstrated by (Pham et al., 2024), even minor LiDAR perturbations (under 2 cm) can induce substantial deviations in 3D obstacle detection accuracy (up to 23.5%) while causing error propagation across downstream ADS components, ultimately compromising system safety. In the domain of 3D point cloud robustness evaluation, (Sun et al., 2022) established ModelNet40-C as a benchmark framework incorporating 15 realistic corruption types, each quantized across five severity levels (i.e., noise intensities). Their analysis revealed a pronounced performance degradation in state-of-the-art models when processing corrupted inputs compared to clean ModelNet40 (Zhirong Wu et al., 2015) data, highlighting critical vulnerabilities in current perception systems.

While Sun et al. (2022) provided comprehensive coverage of various corruption types, our approach focuses specifically on enhancing the modeling of the two most prevalent noise patterns: Occlusion and Background interference (Table 1). For occlusion effects, we depart from their view-angle-based severity scaling in favor of a more physically-grounded approach that modulates the number of blocked LiDAR rays. Regarding background noise, our implementation advances beyond random point generation by incorporating the sensor's spherical coordinate system (following (Wang et al., 2024)) to ensure noise points adhere to the LiDAR's native ray distribution pattern. Furthermore, we constrain noise point generation to valid detection ranges, preventing artificial occlusion by foreground objects. Crucially, our framework establishes explicit mappings between corruption intensity levels and measurable sensor/scenario parameters, enabling more physically meaningful robustness evaluation.

Similarly, (C. Zhang et al., 2024) introduced a LiDAR quality metric that detects environmental anomalies like rain and fog by analyzing spatial and intensity distributions in point clouds. Their method effectively identifies common artifacts such as random noise and intensity distortions across different LiDAR systems.

Li et al. (2021) proposed FLAT (Fool LiDAR with Adversarial Trajectory), a white-box attack that manipulates vehicle trajectories to deceive detection models. While requiring internal model knowledge, FLAT achieves significant disruption - reducing precision by ~70% with just 10cm perturbations and nearly eliminating detection when attacking full trajectories. Similarly, (Bai et al., 2025) developed AdvGLOW, which conducts covert adversarial attacks against autonomous driving perception modules using a Glow-based reversible neural network. By optimizing a combined global-local loss, AdvGLOW introduces reversible data transformations to embed adversarial features. This approach achieves state-of-the-art attack success rates while maintaining human-imperceptible modifications.

Wang et al. (2024) introduced a spoofing method that creates realistic adversarial obstacles by strategically placing sparse point clouds within a LiDAR's detection range, carefully maintaining visibility constraints.

Table 1. Noise/attack type synopsis

Noise type	Physical cause	Implementation	Severity levels
Occlusion	Diffuse reflection/absorption from metallic surfaces causing signal loss	Full-scale saturation of planar ray bundles	S1 = 2 planar rays S2 = 3 planar rays S3 = 4 planar rays S4 = 5 planar rays S5 = 6 planar rays
Background	Atmospheric backscatter (rain, particulates) introducing spurious returns	Random point injection at defined density ratios (new / original points)	S1 = 2.22% density S2 = 2.50% density S3 = 2.86% density S4 = 3.33% density S5 = 5.00% density
Zero-range Background (ZRB)	Near-field obstructions (heavy precipitation, sensor fouling)	Zero-range assignment to specified planar rays	S1 = 1 planar ray S2 = 2 planar rays S3 = 3 planar rays S4 = 4 planar rays S5 = 5 planar rays
Electromagnetic Interference (EMI)	RF interference distorting sensor electronics	Aliased sinusoidal range modulation, severity-dependent amplitude (RF carrier ≈615 MHz)	S1 = 0.5 m amplitude S2 = 1.0 m amplitude S3 = 1.5 m amplitude S4 = 2.0 m amplitude S5 = 2.5 m amplitude

Their approach, which shares similarities with our Background (minor translations) and ZRB (major frontal displacements) noise models, achieved an 89% success rate against Point-Voxel Region-based Convolution Neural Networks (PV-RCNN) detectors (Shi et al., 2020) using just 20 carefully positioned points. In a more sophisticated approach, (He et al., 2023) developed a generative adversarial network (GAN)-based framework that learns to generate adversarial point clouds by simultaneously optimizing adversarial, classification, and reconstruction losses. Their method incorporates dynamic outlier removal to ensure both attack effectiveness and geometric plausibility, demonstrating state-of-the-art performance on ModelNet benchmarks while preserving the structural integrity of point clouds.

Yu et al. (2024) developed an attribution-based method that strategically perturbs critical scanline points to mimic natural disturbances like snow or sandstorms - similar to our Background noise implementation. Meanwhile, (Yang et al., 2024) approached the challenge from a defensive perspective, proposing a CARLA-based (Dosovitskiy et al., 2017) simulation framework that uses a custom Spray Emitter to generate realistic rain-like noise patterns for data augmentation. Their work demonstrates how controlled injection of Background-type noise (simulating raindrops as random points) can improve detector robustness in adverse weather conditions.

Jin et al. (2024) introduced novel cross-modality signal injection attacks to LiDARs by using intentional Electromagnetic Interference. The method induces four types of anomalies: Points Interference, Points Removal, LiDAR Power off, and Points Injection. The researchers validated these anomalies on five LiDAR systems (VLP-16, RD-16, RS-Bpearl, RS-M1 and RS-M1P), demonstrating that the attack remains effective against moving vehicles with success rates exceeding 80% and at distances up to 5.5 meters.

Summarizing, the reviewed studies demonstrate that state-of-the-art 3D perception systems remain vulnerable to both naturally occurring and adversarially crafted point cloud perturbations. These findings motivated our development of a comprehensive toolkit for generating physically grounded LiDAR degradation scenarios that maintain clear interpretability relative to sensor specifications and operational conditions. The resulting benchmark aims at facilitating vulnerability assessment and supporting the development of more robust detection models through physically-grounded testing scenarios.

2.2. Anomaly Detection

Anomaly detection automatically identifies data points or patterns that deviate from expected, normal behavior. Anomaly detection has been extensively studied across various domains (e.g., (Kosuge et al., 2023)), with significant contributions in AD applications. (Baccari et al., 2024) offers a systematic review of AD anomaly detection methods, encompassing both shallow ML and DL approaches. Among these, the OCSVM remains widely adopted due to its conceptual simplicity and effectiveness, but it struggles to capture temporal dependencies as effectively as deep learning and lacks scalability (Kumar et al., 2021). Its versatility is demonstrated in different applications: urban area classification (Ao et al., 2017), high-dimensional sensor data analysis (Qiao et al., 2023), and travel pattern recognition from multimodal GPS data (Zhu et al., 2021).

A notable implementation by (van Wyk et al., 2020) combines OCSVM with an Adaptive Extended Kalman Filter (AEKF) to achieve state-of-the-art accuracy and robustness in real-time sensor anomaly detection for AD systems. Their approach carefully tunes hyperparameters (including anomaly thresholds and contamination percentages) using validation sets containing noisy samples. These implementations consistently incorporate noise characteristics (such as outlier proportions) during model training. However, methodological transparency varies across studies, with several works (e.g., (Li et al., 2024), (Mavikumbure et al., 2024)) failing to document their hyperparameter selection processes - including critical parameters like the contamination factor in Isolation Forest algorithms (Liu et al., 2008) - highlighting a need for more standardized reporting practices in the field.

Beyond OCSVM, anomaly detection techniques commonly employ distance-based, density-based, and random partitioning approaches. (Henriksson et al., 2023) demonstrated the effectiveness of Mahalanobis Distance as an OOD scoring method to filter unreliable predictions in AD semantic segmentation tasks. Their results showed significant reductions in misclassification rates, even on previously unseen datasets.

Xu et al. (2022) achieved state-of-the-art performance on datasets with varying density distributions by combining k-nearest neighbors (kNN) preprocessing with local outlier factor (LOF). Notably, LOF requires representative anomalies in the training data to properly tune its contamination hyperparameter. The efficacy of LOF was further validated by (Swana et al., 2024), who showed superior performance over Isolation Forests in detecting mechanical faults - including brush anomalies and stator/rotor inter-turn faults - while addressing class imbalance challenges. In their study, the authors systematically evaluated different contamination values (30% and 50% outliers) to benchmark detection performance.

Recent studies examining vulnerabilities in LiDAR-based perception systems have extended anomaly detection research to DL architectures. For cybersecurity applications, Prathiba et al. (2023) developed a hybrid deep anomaly detection framework for autonomous vehicles operating in 6G networks. Their decentralized approach enables real-time, collaborative anomaly detection across multiple vehicles while achieving superior accuracy compared to existing methods.

Moreover, Sboui et al. (2025) proposed a hybrid unsupervised anomaly detection framework for vehicle LiDAR sensors based on a CNN-BiLSTM VAE (Kingma and Welling, 2022). The model integrates convolutional layers for spatial feature extraction and bidirectional LSTM layers for temporal dependency learning within a mirror-to-mirror (M2M) VAE architecture. Their method achieved 95.1% accuracy and 82.6% F1-score, outperforming traditional baselines such as OCSVM, LOF, and the hybrid MC-CNN-LSTM AE (Karadayı et al., 2020). The evaluation was conducted on preprocessed multidimensional spatio-temporal LiDAR data from the Lyft motion prediction dataset (Houston et al., 2021), where various environmental anomalies (e.g., multipath reflections, dynamic obstacles, and weather interference) were synthetically injected to simulate real-world conditions. Importantly, the authors tuned the anomaly detection threshold using the test set to maximize F1-score, selecting an optimal reconstruction error threshold of 0.47 for their model. Tuning the threshold based on the test set limits the generalization ability of the trained model.

Industrial applications have driven significant innovations in

production quality monitoring. Luo et al. (2025) advanced beyond conventional AE architectures by using a pre-trained convolutional network with multi-head self-attention for adaptive mask generation, followed by an inpainting network. This approach yields substantially different reconstructions for defective images compared to normal samples. Similarly, Cheng et al. (2025) proposed a novel multi-class point cloud anomaly detection method employing global-local feature matching. Their three-stage process effectively separates easily confusable data before isolating anomalies, demonstrating improved performance on complex industrial inspection tasks. Chaves-Tibaduiza et al. (2024) introduced BERTAD, an encoder-only Transformer model based on BERT, to investigate the feasibility of using such architectures for anomaly detection in IoT network traffic. Trained and evaluated on the MQTT dataset, BERTAD achieved 99.89% accuracy, 98.78% precision, and a Matthews Correlation Coefficient (MCC) of 78.55%, outperforming ML and DL models. The work demonstrates that encoder-only architectures can effectively capture contextual dependencies in high-dimensional data, offering a strong alternative to recurrent or convolutional models for IoT anomaly detection.

2.3. Remarks

The literature review reveals that most state-of-the-art methods explicitly incorporate knowledge about detectable anomalies during model development. This is reasonable, given that many noise patterns and attack vectors are not entirely unforeseen. However, this practice highlights two significant research gaps:

1) Methodological comparison: the field lacks systematic quantitative comparisons between knowledge-based approaches and OOD-unaware novelty detection methods. Such evaluation is crucial for understanding the fundamental detection capabilities of different paradigms.
2) Safety-critical innovation: OOD-unaware novelty detection is still understudied despite its critical importance for autonomous system safety, particularly when encountering truly unforeseen anomaly types that evade predefined detection patterns.

We argue that these gaps represent key limitations in the state-of-the-art. By addressing them, we intend to advance the field toward more robust, generalizable solutions for safety-critical applications.

3 Methodology

This study's methodology addresses the two research gaps through a comprehensive evaluation framework. To bridge the methodological comparison gap, we conduct quantitative comparisons between knowledge-based and pure novelty detection approaches across carefully designed benchmarks. Section 3.1 (Noise/attack types) presents the methodology to systematically evaluate four distinct perturbation typologies (Occlusion, Background, ZRB, EMI) across five severity levels. To address the safety-critical innovation gap, we emphasize OOD-unaware novelty detection by exclusively using clean ID data for model development. Section 3.2 (OOD Input detector models) proposes the new encoder-only Transformer and compares it with the state-of-the-art ATUAD (Yu et al., 2025) and ATF-UAD (Fan et al., 2023). Finally, Section 3.3 (DRL agent data collection) introduces the noise injection protocol and testing environment, designed to validate detection robustness against unforeseen anomalies while maintaining experimental reproducibility.

3.1. Noise/Attack Types

Our evaluation framework incorporates four noise/attack typologies(Occlusion, Background, ZRB, and EMI), each one systematically parameterized across 5 graduated severity levels (Table 1).

Occlusion occurs when LiDAR beams are absorbed or diffusely reflected, creating data voids in the point cloud, with severity ranging from 2 affected planar rays (s1) to 6 rays (s5). Background noise introduces spurious points (e.g., from raindrops/reflections) with a proportion with respect to the total from 2.22% (s1) to 5% (s5), while its specialized variant, Zero-Range Background (ZRB), simulates near-sensor interference (<2m) typical of heavy precipitation (Yang et al., 2025; Zhao et al., 2021), affecting 1 (S1) to 5 (S5) planar rays. For all three point-cloud-modifying perturbations (Occlusion, Background, ZRB), the affected rays are randomized at each frame to represent transient, spatially varying disturbances and to avoid imposing an artificial temporal pattern. This modeling choice is consistent with disturbances characterized by independently varying returns, such as individual raindrop-induced reflections, while persistent or temporally correlated faults are outside the scope of the present corruption model. EMI is instead modeled as a phase-randomized, aliased sinusoidal modulation of the measured LiDAR ranges. The approximately 615 MHz value denotes the assumed RF carrier responsible for the interference (Liu et al., 2024), while the modulation amplitude increases from 0.5 m at S1 to 2.5 m at S5. This taxonomy spans intrinsic sensor interference (EMI: least invasive) to extrinsic disturbance (ZRB: most invasive) and provides controlled severity levels for evaluating a range of operational and adversarial conditions. Overall, the proposed corruption models constitute controlled, physically grounded approximations of selected LiDAR disturbances rather than exhaustive reproductions of real sensor-failure dynamics.

Fig. 3 illustrates the geometrical degradation imposed on the DRL agent's input frame (system architecture detailed in Section 3.3) by four distinct noise modalities. These visualizations highlight the spatial consequences of the attacks on object perception: (1) Occlusion creates isolated spikes at maximum range by affecting rays through beam diffusion/absorption; (2) Background noise introduces sparse random-range false returns; (3) ZRB generates abrupt zero-range drops at random positions, simulating near-field obstructions; while (4) EMI manifests as a smooth, phase-randomized range modulation superimposed on the true signal.

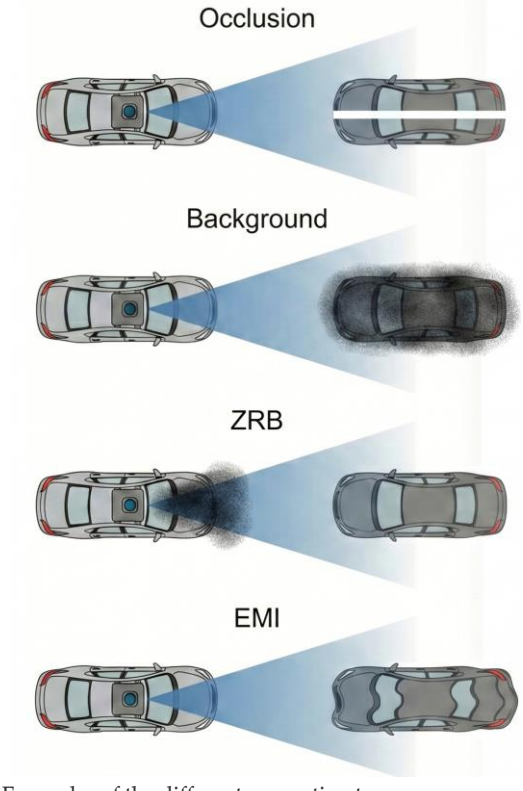

Fig. 3. Examples of the different corruption types.

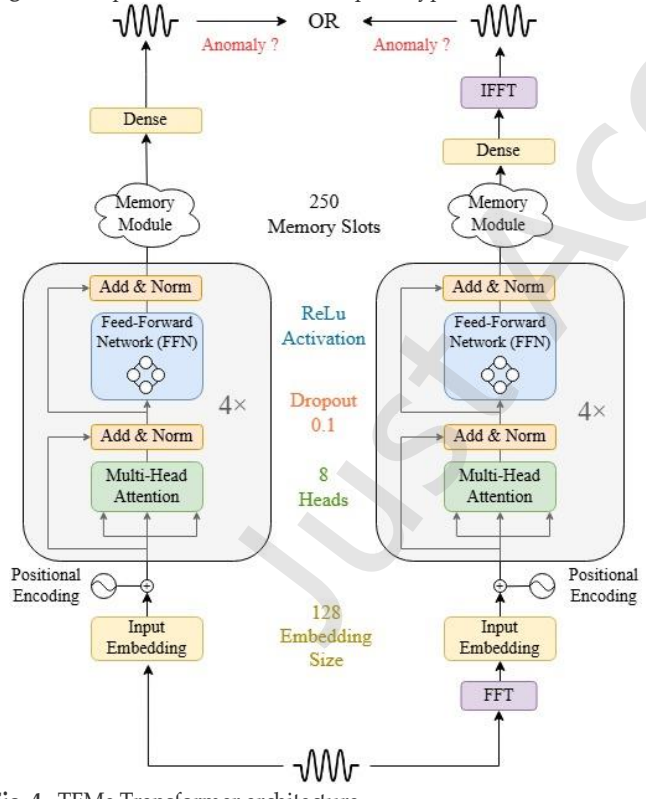

Fig. 4. TFMe Transformer architecture.

3.2. OOD Input Detector

To detect OOD input, we propose the Time-Frequency-Memory enhanced (TFMe) Transformer (Fig. 4). Following established practice in the literature (e.g., (Chaves-Tibaduiza et al., 2024)), we adopt an encoder-only architecture, as the objective is not to generate an output sequence through a decoder, but to extract discriminative features for anomaly detection (Chaves-Tibaduiza et al., 2024). In addition to a branch dedicated to direct time-series processing, the model incorporates a parallel branch that operates on the frequency-domain representation of the signal. This design captures multi-scale, time and frequency contextual dependencies, thereby improving ID representation. For example, this contributes to alleviating the impact of occlusions that tend to impair the time-domain branch. Memory modules were added to the two domains to prevent the over-generalization problem, since some anomalies may share common compositional patterns with ID training data (Gong et al., 2019).

The proposed TFMe differs from the state-of-the-art for design. ATUAD (Yu et al., 2025) is a single-domain, memory-less encoder–decoder architecture relying on adversarial training, while ATF-UAD (Fan et al., 2023) reconstructs time and frequency views through an adversarial network without attention-based memory. TFMe, instead, is an original combination of a dual-branch, encoder-only Transformer, per-domain memory modules, a consistency loss aligning the two reconstructions, and a disjunctive (OR-gate) decision rule.

The model first creates linear embeddings, which are augmented with fixed (i.e., non-learnable) sinusoidal positional encodings. As the input data is sequential, Positional Encoding (PE) is critical for preserving the temporal order of timesteps. Embeddings (of size 128) are then processed by a stack of multi-head self-attention (MHSA) layers via a dual-branch structure. The frequency-domain branch processes inputs transformed by the Fast Fourier Transform (FFT), while the time-domain branch operates on the raw signal embeddings.

To enforce sparsity in the memory addressing weights, two types of weight regulators are applied: entropy loss and hard shrinkage (Gong et al., 2019). The importance of their combination lies in preventing the model from reconstructing anomalies by "cheating", using a complex combination of many memory items. The entropy loss is an additional training loss term that penalizes the entropy of the weights (instead of L1 regularization). Hard shrinkage physically truncates small weights by zeroing any weight below the hard shrinkage threshold (λ). We determined the optimal threshold by evaluating {1/N, 3/N} (with N=250, representing the number of memory slots), ultimately adopting $\lambda = 1/N$.

Following the memory module, the latent representation is passed through a final Dense (Fully Connected) subnetwork. This network acts as a decoder head, projecting the dimensional embedding vectors back to the original feature space through dense layers (1 hidden layer for the time branch and 2 for frequency). In the time-domain branch, this layer directly outputs the reconstructed signal values. In the frequency-domain branch, the layer outputs the predicted frequency spectrum, which is reconverted to the time domain via the Inverse Fast Fourier Transform (IFFT).

Given that the time and frequency domains represent the same physical reality for ID data, the reconstructions from both branches are expected to be consistent. Assigning disproportionate weights to one branch would impose an explicit prior belief regarding the nature of expected anomalies, thereby biasing the model toward specific temporal or spectral deviations (Fan et al., 2023). Furthermore, combining both losses introduces a significant drawback, since the domain with the higher threshold tends to dominate the detection process. Consequently, an anomaly arising in the non-dominant domain may be classified as an ID rather than OOD. To mitigate this issue and maximize sensitivity, we adopt a disjunctive logic (OR gate)

for the final decision layer. A sample is classified as an anomaly if it is flagged by either the time or frequency domain branch.

The loss function is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{entropy} + \lambda_{con} \quad (2)$$

The total loss function sums the reconstruction loss $\mathcal{L}_{rec}$ (MSE between the original and the reconstructed signal), the entropy loss $\mathcal{L}_{entropy}$ (for the memory slots) and the consistency loss $\mathcal{L}_{con}$ (for dual-branch alignment (Zhou et al., 2022)). The $\lambda_{entropy}$ weighting hyperparameter was set to 0.0002, and λ_{con} was set to 0.1 after testing {0, 0.01, 0.1, 1} values, as suggested by (Gong et al., 2019).

The candidate and optimal hyperparameter values for the TFMe architecture and training process are reported in Table 2. The optimal configuration uses 4 encoder layers per domain, an embedding dimension (d_{model}) of 128, and positional encoding with a maximum length of 10. Each encoder layer incorporates 8 attention heads, followed by a Feed-Forward Network (FFN) with a dimension of 512 and ReLU activation. The memory size was set to 250. The reconstruction head consists of two dense layers: a hidden layer of size $2d_{model}$, with Leaky ReLU activation, followed by a final output layer responsible for reconstructing the original signal. Optimization was performed using Adam (lr=1e-4, dropout=0.1, batch_size=256, epochs=500). The proposed time-domain Transformer (TMe) employs the identical architecture, but with no frequency-domain branch.

Table 2. Tested hyperparameter values for the TFMe

Model	Hyperparameter	Tested values	Best values
Transformer	Number of layers	1-5	4
	D_model	32, 64, 128, 256	128
	Optimizers	Adam, AdamW	Adam
	Dimension of FFN	64, 128, 256, 512, 1024	512
	Number of heads	2, 4, 8, 16	8
	Activation	ReLU, GELU	ReLU
	Learning rate	1e-3, 1e-4	1e-4
	Dropout	0.1-0.3	0.1
	Memory Slots	50-500	250
	Dense Layers	1-3	2
	Dense Activation Function	ReLU, Leaky ReLU	Leaky ReLU

We compare our model with ATUAD (Yu et al., 2025), which is a Transformer-based encoder-decoder architecture with an adversarial training framework to capture long-distance contextual information while enhancing sensitivity to mild anomalies. The architecture consists of a shared encoder and two independent decoder branches (to generate the attack) that undergo a two-phase training process: initially minimizing reconstruction errors and subsequently engaging in an adversarial loop where one transformer block attempts to deceive the other to amplify reconstruction deviations for abnormal data.

The model was optimized using AdamW, trained for 100 epochs with mini-batches of 128 samples and a learning rate of 1e-3. To prevent overfitting, an early stopping patience of 10 was used and a dropout of 0.1. The model uses 128 as an embedding size, following empirical tuning.

On the other hand, Fan et al. (2023) introduced the Adversarial Time-Frequency Reconstruction Network for Unsupervised Anomaly Detection (ATF-UAD), a framework designed to reconstruct input sequences by leveraging both time and frequency domain perspectives. The architecture features a time reconstructor that employs parity sampling, attention mechanisms, and Graph Convolutional Networks (GCNs) to weaken neighbor dependencies and dilute the influence of anomalies. Simultaneously, a frequency reconstructor utilizes Fourier transforms to extract and reconstruct phase and amplitude information, effectively filtering anomalies with complex frequency characteristics. These components are integrated through a dual-branch adversarial learning mechanism that minimizes reconstruction errors while maximizing residual outliers to highlight anomalies.

The model was optimized using Adam with a weight decay of 1e-5, trained for 500 epochs with mini-batches of 128 samples and a learning rate of 1e-4. The ATF-UAD model needs configuring two additional hyperparameters: σ (which controls the weights in dual-view adversarial learning) and β (which balances the time and frequency sequences in the final reconstruction). Both were set to 0.5. We adopted this balanced approach to ensure that equal importance is assigned to time and frequency domains.

In all our experiments, we train the models on ID data only. In training, the models are optimized using their loss functions (i.e., reconstruction, entropy, and consistency losses). During inference, samples are flagged as OOD if their reconstruction error exceeds a predefined threshold. In the OOD-aware setting, the threshold is selected by maximizing the F1-score on validation data containing labelled OOD samples. In the OOD-unaware setting, the threshold is calibrated exclusively on ID validation reconstruction errors to control the false-positive rate (FPR). This design choice aligns with automotive safety priorities, as excessive false alarms would undermine user trust by generating unnecessary TORs in operational environments.

The validation set contains only ID samples in the OOD-unaware (pure novelty) approach, while it contains also OOD data in the OOD-aware method. For the OOD-unaware approach, the decision thresholds are calibrated exclusively from the reconstruction-error distributions obtained on the ID validation set. We considered both the empirical 99th percentile and the three-sigma rule, defined as the mean plus three standard deviations. The two criteria yielded very similar thresholds: 0.0074 and 0.0076, respectively, for the time-domain branch, and 0.0062 and 0.0064 for the frequency-domain branch. Moreover, the resulting detection metrics were identical at the reported precision. We therefore adopt the empirical 99th percentile as the primary calibration rule because it does not require a distributional assumption, while retaining the three-sigma criterion (which assumes an approximately normal reconstruction-error distribution) as a sanity check. The close agreement between the two methods indicates that, for the considered validation data, the detector performance is not sensitive to the specific statistical calibration rule.

For reproducibility, all models were trained with a fixed random seed (seed = 42). To verify seed-independency, each model was additionally retrained with 5 independent seeds, with the decision threshold re-derived for each run on the corresponding ID validation set. The validation FPR varied by less than 0.2%. Models were developed under

identical experimental conditions: the same episode-level train-validation-test splits (Section 3.3), the same pre-processing, comparable hyperparameter search budgets (Table 2 for TFMe/TMe; and the hyperparameter grids explored for ATUAD and ATF-UAD follow the respective original papers). So, the observed differences can be attributed to the detection paradigms rather than to the experimental set-up. The variability of the reported metrics was estimated at evaluation time through 20 Monte Carlo folds of 500 samples each (Table 8). Training required approximately 48 min (TFMe), 30 min (TMe), 48 min (ATUAD), and 4 min (ATF-UAD) on an NVIDIA RTX 4060 GPU. Model sizes and inference costs are reported in Section 4.5.

As for pre-processing, we only employed the MinMax scaler for all models. Initially, we explored additional pre-processing techniques to evaluate whether removing outliers from the ID data could enhance the performance of the subsequent anomaly detection models. In fact, while the FPR=0 constraint is given by the absence of positive samples during training, outlier elimination could lead to more robust overall performance. To this end, we applied three unsupervised learning methods: DBSCAN (Schubert et al., 2017), Gaussian Mixture Models (GMM) (Yu and Deng, 2015), and Local Outlier Factor (LOF) (Breunig et al., 2000). However, we ultimately dropped this step, as it consistently led to degraded performance on the validation set. This outcome may be attributed to the absence of actual outliers in the simulated environment, likely due to the idealized nature of the simulated sensors. Greater caution and consideration are warranted when applying such techniques to real-world data.

3.3. DRL Agent Data Collection

To evaluate our OOD detection system, we developed a test scenario using a state-of-the-art DRL agent for low-speed parking maneuvers in CARLA (Lazzaroni et al., 2024), which is the DRL Agent in Fig. 2. The experimental setup features comb-shaped parking spaces with the following procedural generation parameters for each training episode: (1) random selection of a target parking lot, (2) dynamic spawning of 0-20 vehicles in the vicinity to create varying obstacle densities, and (3) controlled initialization of the ego vehicle with a fixed ±5 m lateral offset from the target's y-axis (while ensuring environment boundary constraints) and random 360° orientation. This configuration systematically induces various parking scenarios while maintaining reproducibility through constrained randomization.

The ego vehicle (EV) features a roof-mounted 360° short-range LiDAR sensor with 6° horizontal and 5° vertical angular resolution, operating at 5 Hz. To balance DRL training efficiency and operational accuracy (Berta et al., 2024), the raw point cloud is pre-processed into 61 planar rays representing minimum object distances per horizontal sector, aggregated across vertical angles. This perception system is augmented with seven vehicular state parameters (target distance, speed, acceleration, and angular velocity) per frame. The DRL agent's input stacks five temporally consecutive frames (1s at 5Hz) into a 340-dimensional input vector (i.e., (61 LiDAR + 7 vehicle parameters) × 5 frames). We highlight that the noise injection pipeline (Section 3.1) operates before this dimensionality reduction. LiDARacks corrupts the original CARLA 3D point cloud, and the corrupted cloud is then passed through the same pre-processing stage used by the DRL agent, which retains, for each horizontal angular sector, the minimum obstacle distance aggregated over the vertical field of view. This compact sector-based representation is a deployment-oriented choice for the considered self-parking ADF. Consequently, the present conclusions apply primarily to ADFs employing similar reduced LiDAR inputs and cannot be directly generalized to perception pipelines operating on full 3D point clouds.

The agent architecture comprises a two-layer Multi-Layer Perceptron (MLP) (512 neurons per hidden layer) with 4 output neurons corresponding to discrete driving actions: steering angle, throttle, brake, and reverse. In standalone testing, this configuration achieves 97% successful parking completion, with 2% collisions and 1% timeouts. To avoid collisions in real-world deployment, this DRL core would be integrated within a safety-critical rule-based wrapper that intervenes when obstacle proximity thresholds are exceeded, as standard in automotive ADF architectures. But this approach is vulnerable to noise.

We compiled our dataset from 718 successful parking episodes executed by the trained agent. The 718 episodes were first divided into episode-disjoint training, validation, and test sets comprising 513, 92, and 113 episodes, respectively. Five-frame samples were then generated within each partition, preventing samples from the same episode from appearing in different sets (Table 3). To evaluate OOD detection performance, we generated a balanced test benchmark by adding noise to the original point clouds collected from additional 113 episodes. We employed LiDARacks to generate all four noise/attack types (Occlusion, Background, ZRB, EMI) across five severity levels. This produced 10,704 OOD samples for each noise/severity pair, matching the ID test set size for statistically robust comparison. The noise injection applies perturbations uniformly across each episode's duration.

Table 3. Data and Partitions

Type	Episodes	ID samples	OOD samples
Train	513	50,335	-
Val	92	10,006	10,006 (OOD-aware case only)
Test	113	10,704	10,704
Total	718	71,045	10,704 (+10,006)

Restricting the ID dataset to successful episodes provides a representation of nominal operation, but may underrepresent rare yet valid maneuvers and operating states close to failure.

4 Experimental evaluation

This section intends to validate our methodology by systematically evaluating anomaly detection performance across two key dimensions: OOD-aware tuning, exploiting knowledge on specific noise types (Section 4.1.1); and OOD-unaware tuning, assessing pure novelty detection (4.1.2). Sections 4.1.3-4 deepen the novelty detection performance through, respectively, a robustness evaluation, and an explainability analysis, via SHAP. Section 4.2 offers a performance comparison on real-world sensor data. Section 4.3 assesses the generalization of the detectors to an unseen car park layout. Moving to the automotive system architecture, the TOR management study in 4.4 verifies the design's robustness across multiple operational scenarios. Finally, a deployability analysis in 4.5 demonstrates real-time feasibility on embedded platforms. The overall goal is to empirically bridge the methodological comparison gap by quantifying detection trade-offs (Sections 4.1-3) and address the safety-critical innovation gap by

Table 4. Test results for OOD-aware ATUAD (Yu et al., 2025)

	EMI @S1				EMI @S5				OCC @S1				OCC @S5			
	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR
EMI @S1	90%	92%	91%	10%	91%	99%	95%	10%	87%	67%	76%	10%	90%	93%	91%	10%
EMI @S5	98%	47%	64%	1%	99%	97%	98%	1%	97%	21%	35%	1%	99%	60%	75%	1%
OCC @S1	78%	98%	87%	28%	78%	99%	87%	28%	75%	84%	79%	28%	78%	98%	87%	28%
OCC @S5	91%	91%	91%	10%	91%	99%	95%	10%	87%	66%	75%	10%	91%	93%	92%	10%

Table 5. Test Results For OOD-aware ATF-UAD (Fan et al., 2023)

	EMI @S1				EMI @S5				OCC @S1				OCC @S5			
	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR
EMI @S1	89%	91%	90%	12%	89%	95%	92%	12%	88%	85%	86%	12%	89%	98%	93%	12%
EMI @S5	95%	74%	83%	4%	96%	94%	95%	4%	94%	73%	82%	4%	96%	95%	95%	4%
OCC @S1	86%	95%	90%	15%	86%	97%	91%	15%	85%	89%	87%	15%	87%	99%	93%	15%
OCC @S5	93%	81%	87%	6%	94%	94%	94%	6%	93%	78%	85%	6%	94%	96%	95%	6%

Table 6. Test results for OOD-aware TMe Encoder-Only Transformer

	EMI @S1				EMI @S5				OCC @S1				OCC @S5			
	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR
EMI @S1	88%	98%	93%	13%	88%	99%	93%	13%	88%	93%	90%	13%	88%	99%	93%	13%
EMI @S5	99%	69%	81%	1%	99%	98%	98%	1%	99%	65%	78%	1%	99%	91%	95%	1%
OCC @S1	89%	98%	93%	13%	89%	99%	94%	13%	88%	93%	90%	13%	89%	99%	94%	13%
OCC @S5	97%	88%	92%	3%	97%	99%	98%	3%	97%	77%	86%	3%	97%	97%	97%	3%

Table 7. Test results for OOD-aware TFMe Encoder-Only Transformer

	EMI @S1				EMI @S5				OCC @S1				OCC @S5			
	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR	Prec.	Rec.	F1	FPR
EMI @S1	92%	97%	94%	9%	92%	99%	95%	9%	91%	92%	91%	9%	92%	99%	95%	9%
EMI @S5	99%	71%	83%	1%	99%	98%	98%	1%	99%	64%	78%	1%	99%	94%	96%	1%
OCC @S1	92%	97%	94%	9%	92%	99%	95%	9%	91%	92%	91%	9%	92%	99%	95%	9%
OCC @S5	97%	85%	91%	2%	98%	99%	98%	2%	97%	79%	87%	2%	98%	97%	97%	2%

proving viability of pure novelty detection in autonomous systems (Sections 4.4-5).

4.1. OOD Input Detector

4.1.1. OOD-Aware Tuning

As a preliminary experiment, we evaluated the generalization capability of models tuned with the inclusion of OOD data. While it is reasonable to expect that detectors tuned on specific types of noise or attack may exhibit a bias toward those patterns, we are not aware of any studies in the literature that provide a comprehensive and quantitative analysis of this phenomenon.

For each noise type, we collected 5,000 samples, considering both the lowest (S1) and highest (S5) severity levels of EMI and Occlusion noise. Each one of the anomaly detection model types (ATUAD, ATF-UAD, TMe, TFMe) was individually tuned for every specific (noise type, severity level) combination. Here the decision threshold is a hyperparameter, selected by F1-score maximisation on the validationset, which contains OOD samples of that type and severity. To evaluate generalization, each trained model was also tested across all other combinations.

In the resulting tables (4-7, one for each model type), each row corresponds to the noise type and severity level used in validation, while each column represents the type and severity level used for testing. Cells shaded in gray indicate homologous testing scenarios, where the model was evaluated on the same (type, severity) pair it was trained on. Overall, the proposed dual-domain Transformer (TFMe) consistently outperforms the other models in maximizing the F1-score, and in minimizing FPR.

The models' generalization capability is highly sensitive to the severity level used in validation. When tuned on EMI @S5, the TFMe model achieved its lowest FPR (1%) across all noise types and intensities. Although the model generalized effectively to OCC@S5 (96% F1-score compared to 97%), it exhibited performance degradation at lower severity levels (S1), specifically showing a 26% decrease in recall for EMI@S1 and 28% for OCC@S1. Conversely, when the same tuning approach was applied to OCC@S5, the model exhibited recall decreases of 12% for EMI@S1 and 13% for OCC@S1. We attribute this performance degradation to the use of a relatively high threshold on a noise type that closely resembles ID data, resulting in a significant loss of sensitivity.

When tuned on EMI and OCC @S1, TFMe applies a similar threshold across them, yielding acceptable generalization performance. This outcome is expected, as such noise closely resembles ID data, thereby making models sensitive. Under this tuning, the models compensate by adopting lower reconstruction error thresholds, resulting in higher false positive rates to minimize false negatives. Consequently, while the TFMe maintained comparable performance on other noise types, it exhibited a significantly higher FPR (9%) compared to other combinations (1% for EMI@S5 and 2% for OCC@S5).

This trend is consistent across the state-of-the-art models, as reflected in Tables 4 and 5. For instance, tuning on EMI@S5 resulted in a recall

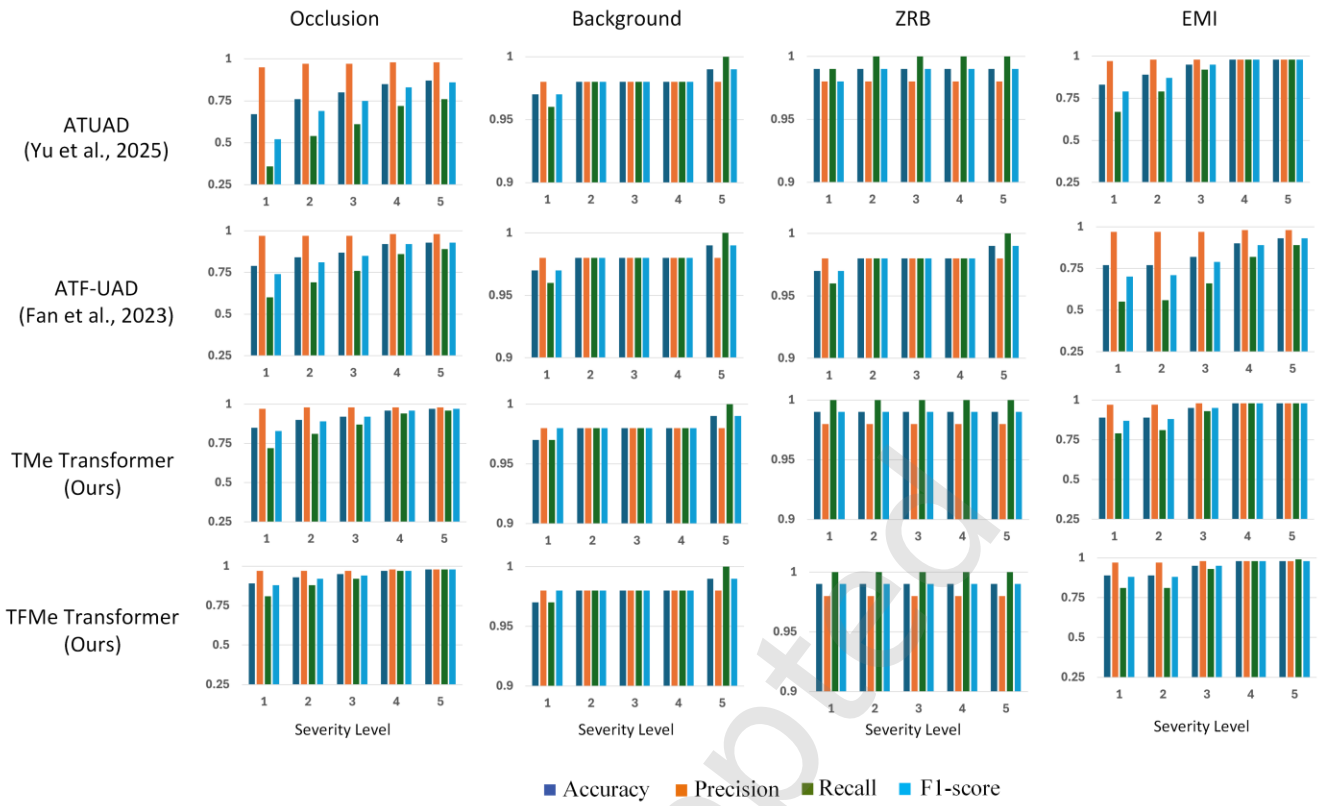

Fig. 5. Models' performance by noise type and severity in the OOD-unaware experiment.

drop of 45% and 17%, for ATUAD and ATF-UAD respectively, when tested on the same noise type but at lower severity level (EMI@S1). Furthermore, tuning models on OCC@S1 introduced the highest FPRs among all noises, reaching 28% for ATUAD and 15% for ATF-UAD.

Overall, these findings highlight a critical limitation: models tuned on targeted, especially invasive, noise types tend to lack generalization across various OOD disturbances. This poses a significant concern for safety-critical real-world applications, where the nature and severity of anomalies are often unpredictable. Ensuring robust generalization is therefore essential to maintain reliability and safety in dynamic operational environments.

4.1.2. OOD-Unaware Tuning

Results for the OOD-unaware novelty detection approach are reported in Fig. 5. ATUAD and ATF-UAD perform competitively (high F1-score) across all severity levels for Background and ZRB noise.

The TMe model consistently outperforms both these baselines across all severity levels of Occlusion, which is a critical vulnerability for 3D point cloud recognition models (Sun et al., 2022). For instance, TMe achieved F1-scores ranging from 83% to 97% on Occlusion severities 1-5, compared to 52-86% for ATUAD and 74-93% for ATF-UAD. Moreover, regarding EMI at severity levels 1 and 2, TMe achieved F1-scores of 87% and 90%, respectively, outperforming both ATUAD (79% and 87%) and ATF-UAD (70% and 71%).

This performance gain could be attributed to the memory module, which prevents the confident reconstruction of OOD samples (over-generalization issue (Gong et al., 2019)) by suppressing weights for memory items with low similarity.

Still, TMe misses some OOD detections (particularly: low recall on low-severity Occlusion samples), which spurred us to design a full TFMe model. Specifically, the frequency-domain branch improves the detection by recognizing that the frequency bins generated by anomalies are not present in the ID frequency memory slots.

In general, TFMe consistently demonstrates the strongest anomaly detection capability and the most balanced overall performance (balanced F1-score in Fig. 5).

Comparing the results obtained with the OOD-aware (Tables 4-7) and OOD-unaware (Fig. 5) approaches sheds significant light on the different generalization ability of the two methods. Section 4.1.1 indicates two main issues with the OOD-aware approach: the drop in recall when testing high-severity thresholds on lower levels, as well as the high FPR at low severity. The OOD-unaware model prevents these problems and demonstrates superior generalization across all anomaly types and severity levels. For example, TFMe achieves a steady F1-score of 88% and a precision of 97% at the lowest severity level for both EMI and Occlusion. As the severity increases to level five, the F1-score and precision reach 98%. This confirms that the OOD-unaware approach establishes a robust decision boundary that is effective regardless of the specific intensity or type of the OOD data.

To validate the performance improvements observed across corruption types, we conducted a Wilcoxon Signed Rank Test on paired F1-scores obtained from 20 Monte Carlo folds of 500 samples each. The analysis further reveals distinct robustness profiles against the ATUAD and ATF-UAD baselines. In the Occlusion scenario, TFMe reduced structural data loss impact across all severity levels, improving F1-scores by up to 71.9 percentage points over ATUAD and 18.5 percentage points over ATF-UAD. These improvements are statistically significant and exhibit large effect sizes (Wilcoxon signed-rank, $p = 9.54e\text{-}7$; Rank-biserial correlation $r = 1.00$; Cliff's $\delta = 1.00$). Under EMI noise, TFMe outperformed ATF-UAD at all severities (up

to +42.0 percent point F1-score improvement; $p \leq 4.26e\text{-}5$; $r \geq 0.995$; $\delta \geq 0.91$) and achieved significant early-stage gains over ATUAD at Severity 1 (+20.0 percent point; $p = 9.54e\text{-}7$; $r = 1.00$; $\delta = 1.00$), demonstrating better robustness to initial signal degradation. On the other hand, no statistically significant improvements were observed for the ZRB or Background datasets (p=1.0), since all evaluated models achieved near-perfect performance under this noise type.

Furthermore, we analyzed the standard deviation of the F1-score across folds as an estimator of predictive stability. Table 8 reports the average across the 20 folds of the F1-score averaged over the severity levels. The proposed TFMe model demonstrated high robustness across all categories. For Occlusion and Background noises, TFMe achieved the highest F1-scores (0.8937 and 0.9493, respectively) while simultaneously maintaining the lowest SD compared to the TMe model and other baselines, indicating high accuracy paired with stability. In the case of ZRB, although TFMe exhibited a slight variance (SD=0.0040) compared to the zero-variance results of TMe and ATUAD, the lower bound performance interval (0.9540) still exceeds the maximum average score of the TMe and ATUAD models (0.9500), proving that TFMe outperforms the baselines even when accounting for variability. Similarly, for EMI, TFMe kept the lead in F1-score (0.9144) with a stability profile comparable to TMe (0.9047) and significantly better than the other UAD baselines (0.9081 and 0.7948).

Table 8. Average and standard deviation in parenthesis across the 20 evaluation folds of the F1-score averaged over the five severity levels, per noise type.

Models	Occlusion	Background	ZRB	EMI
TFMe	0.8937 (0.0372)	0.9493 (0.0057)	0.9580 (0.0040)	0.9144 (0.0424)
TMe	0.8627 (0.0567)	0.9447 (0.0063)	0.9500 (0.0000)	0.9047 (0.0421)
ATUAD	0.7034 (0.1150)	0.9408 (0.0069)	0.9500 (0.0000)	0.9081 (0.0447)
ATF-UAD	0.8151 (0.0677)	0.9387 (0.0075)	0.9453 (0.0097)	0.7948 (0.0833)

4.1.3. Robustness Evaluation

To verify that the reported ranking is not biased toward a specific threshold choice, we evaluate the three detectors under four threshold-free criteria, reported in Figs. 6-9. Fig. 6 reports partial ROC curves restricted to FPR ≤ 0.1, Fig. 7 reports precision-recall and Average Precision (AP) values, Fig. 8 reports recall at fixed FPR, and Fig. 9 reports the FPR at fixed recall. We restrict the ROC analysis to the low-FPR region because higher rates would compromise user acceptance. For the same reason we report the partial AUC (pAUC) normalised over FPR ≤ 0.1.

Fig. 6 shows near-ideal separability for all three models under ZRB and Background noise, with pAUC values converging to 1.000 from severity 2 upwards. This behaviour follows from the geometrically abrupt nature of these two perturbations, which produce reconstruction errors well outside the ID distribution. The UAD models have different performance in Occlusion and EMI. Under Occlusion at severity 1, TFMe reaches a pAUC of 0.906, against 0.834 for ATF-UAD and 0.782 for ATUAD, and keeps the lead over the whole severity range (0.984 against 0.957 and 0.943 @S5). The same ranking holds under EMI, where TFMe spans 0.809 at S1 to 0.993 @S5, against 0.696-0.960 for ATF-UAD and 0.736-0.984 for ATUAD. The margin is largest at low severities, where the perturbed signal most closely resembles ID data.

Fig. 7 reports the precision-recall curves and the corresponding Average Precision (AP). ZRB and Background saturate for all detectors (AP ≥ 0.99), while Occlusion and EMI discriminate between them, most clearly at severity 1. There, TFMe attains an AP of 0.97 under Occlusion and 0.94 under EMI, against 0.94 and 0.89 for ATF-UAD and 0.91 and 0.89 for ATUAD. The ordering matches that of the partial ROC analysis, indicating that the advantage of TFMe does not depend on the region of the operating characteristic selected for the comparison.

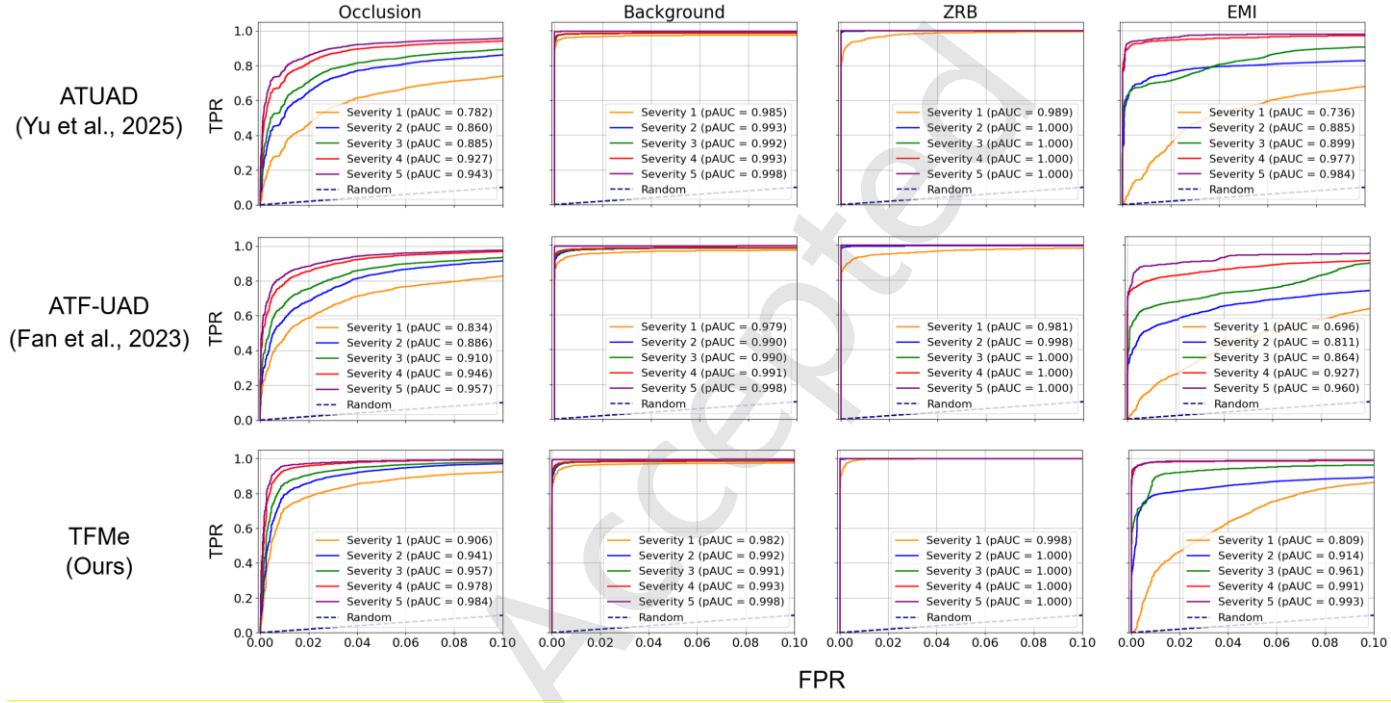

Fig. 6. Partial ROC curves (FPR ≤ 0.1) and pAUC values for the compared models under the four noise types.

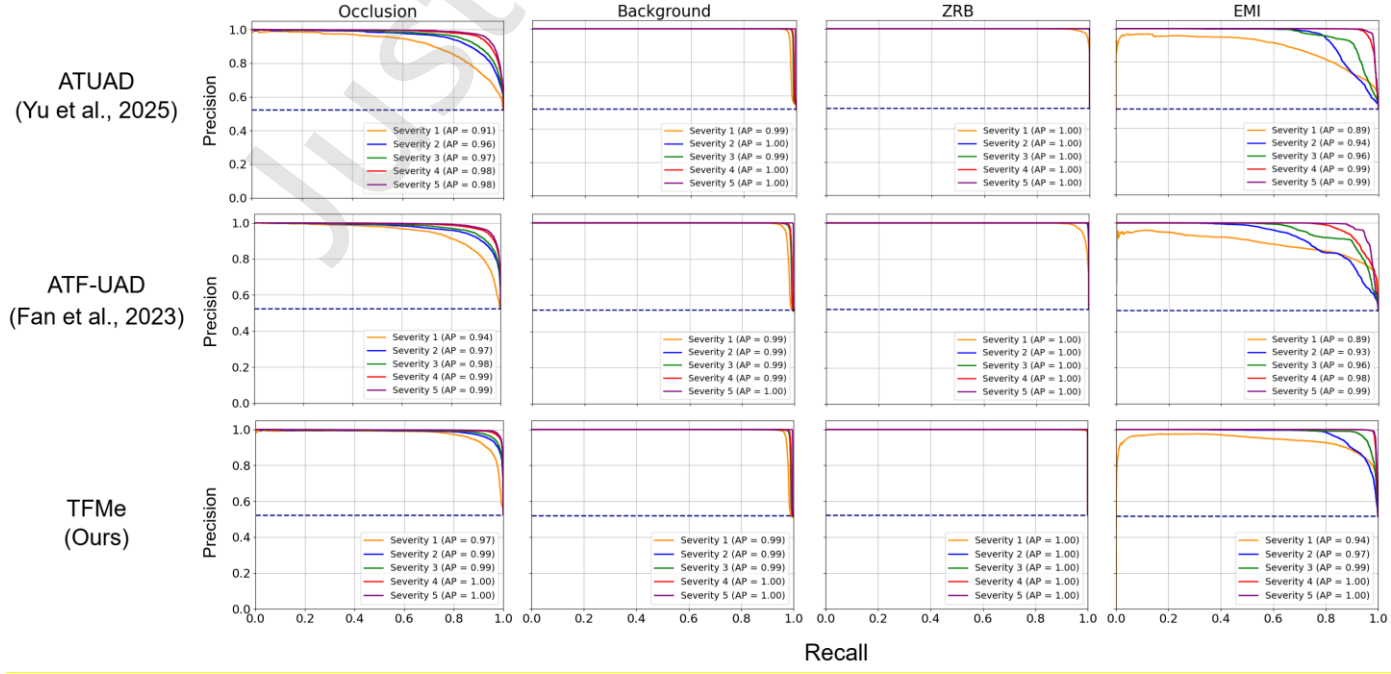

Fig. 7. Precision-recall curves and Average Precision (AP) values for the compared models under the four noise types.

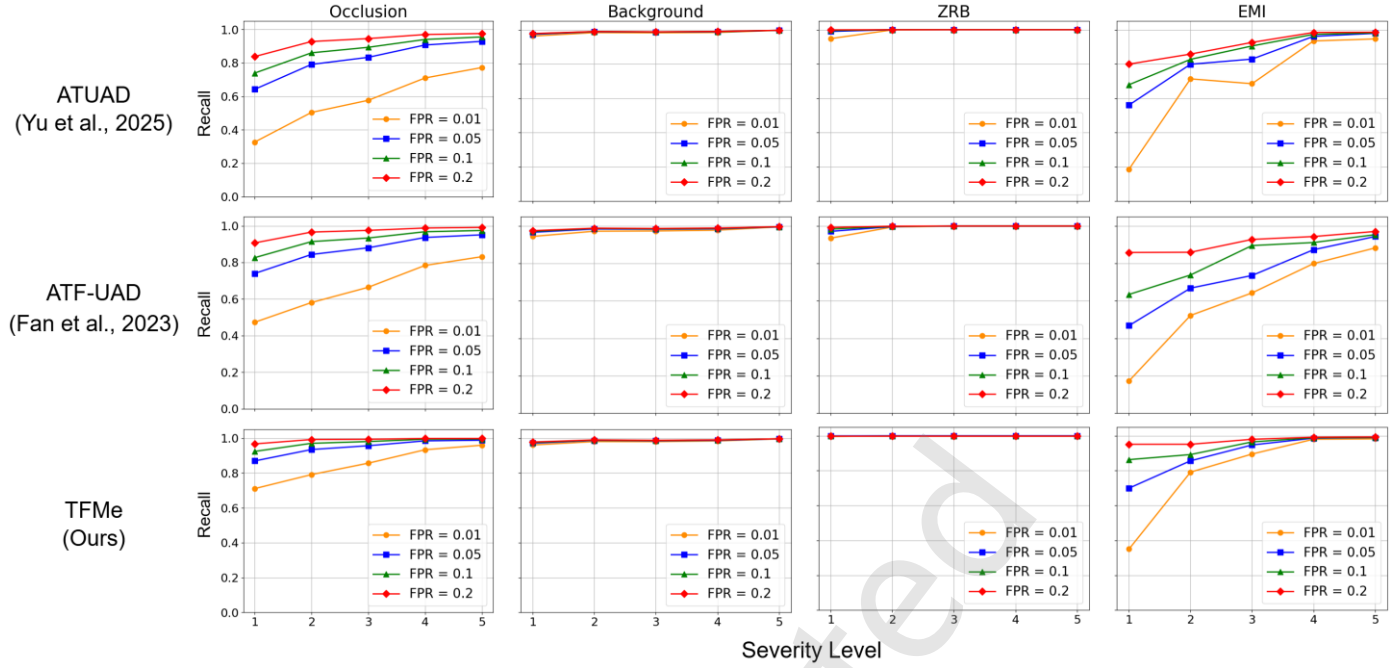

Fig. 8. Recall under a fixed false-alarm budget for the compared models under the four noise types.

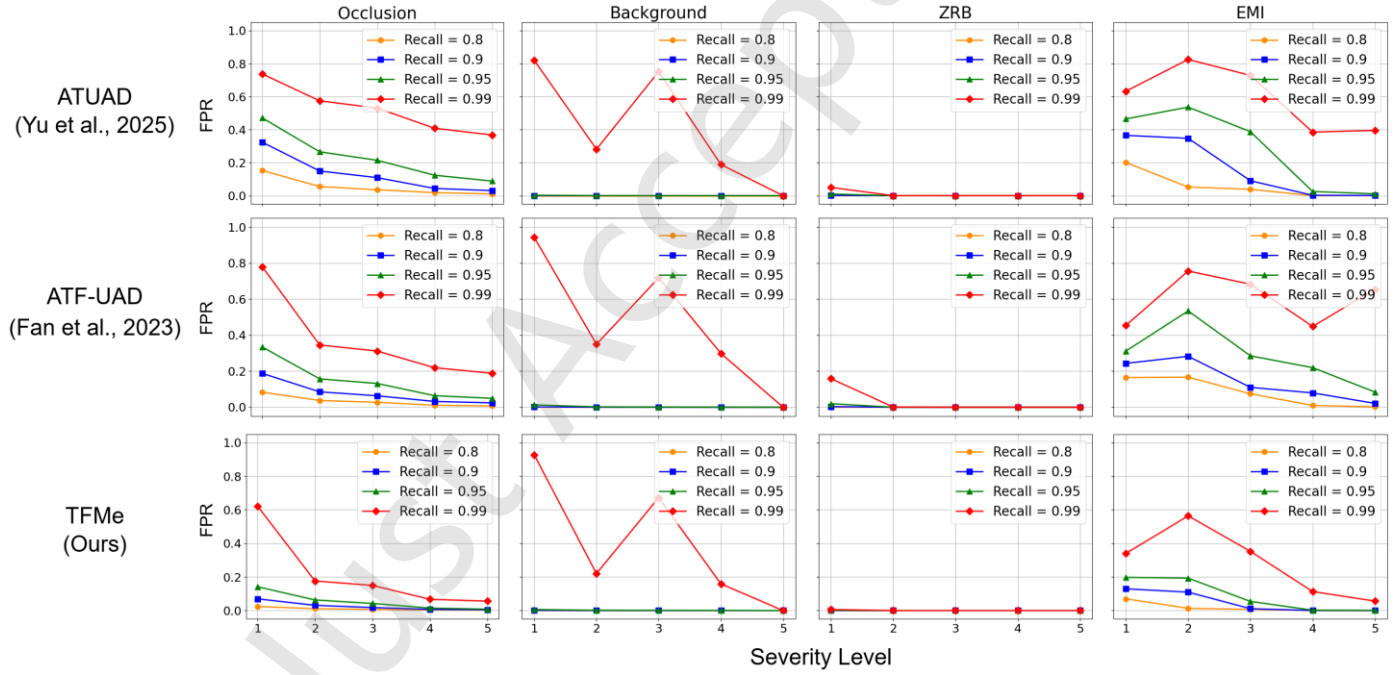

Fig. 9. FPR required to guarantee a prescribed recall for the compared models under the four noise types.

Fig. 8 reports the recall achievable under a fixed false-alarm budget, which reflects the of minimising false positives to preserve user trust. Under Background and ZRB noise, all models exceed 0.95 recall at every budget and severity, matching the saturated F1-scores in Table 8. Under Occlusion, the models diverge sharply at the tightest budget. At FPR = 0.01 and severity 1, TFMe detects about 70% of the anomalous samples, against 46% for ATF-UAD and 32% for ATUAD. TFMe leads at every severity and reaches near-unit recall earlier than the baselines under all budgets, with the largest margin at the tightest budgets and lowest severities. Under EMI at severity 1, all models degrade at FPR = 0.01, but the ranking holds, and TFMe reaches near-unit recall by severity 4 under all budgets. The narrow spread among TFMe's four budget curves points to a sharper separation between the ID and OOD score distributions. This threshold-independency matters most in the OOD-unaware setting, where no anomalous samples are available to refine the threshold, so robustness to threshold placement becomes a prerequisite for reliable deployment.

Fig. 9 shows the false-alarm cost of guaranteeing a prescribed recall. Under ZRB, all recall targets up to 0.99 are met at essentially zero FPR across every severity. Under Occlusion at severity 1, TFMe reaches the 0.95 recall target at an FPR of about 0.13, against roughly 0.33 for ATF-UAD and 0.46 for ATUAD, a 2.5-to-3.5x lower false-alarm cost that persists through severity 5. EMI shows the same

pattern in which TFMe needs about 0.20 FPR for 0.95 recall at severity 1, versus 0.30 and 0.46 for the baselines. The non-monotone profile in the Background column at the 0.99 target arises because nearly all anomalies are already detected at negligible FPR. The last percentile of samples cannot be detected by any of the three models. Overall, these threshold-independent results support the importance of OOD-unaware novelty detection to produce a robust decision boundary across anomaly types and intensities.

4.1.4. Explainability

While deep learning models often operate as black boxes, SHAP (Dabbous et al., 2024; Trivedi et al., 2024) interprets their decisions by quantifying each feature's contribution. In our study, we employed the Gradient SHAP explainer (1,000 training samples, 500 test samples) to analyze the Transformer's temporal attribution profile across the five consecutive LiDAR frames (1s observation window) used by the DRL agent. Table 9 presents the average absolute SHAP values per frame, showing that the attribution profile is distributed across all five frames, with modestly higher average attribution for the first and most recent observations. This suggests that the boundary frames may have somewhat greater influence, with the first frame potentially providing an initial contextual reference.

This temporal attribution profile is coherent with both the physical constraints of vehicle dynamics (where recent data best predicts immediate hazards) and the benefits of short-term memory for distinguishing true anomalies from transient noise.

These SHAP explanations support three practical engineering tasks. For debugging, the relatively high attribution assigned to F(0) provides a sanity check on the chronological stacking and indexing of the input frames. A systematic attribution peak at an unexpected position could indicate reversed, shifted, or otherwise misaligned frames in the preprocessing pipeline. For validation, the non-negligible contribution of all five frames suggests that the detector does not rely exclusively on a single-frame shortcut. Although SHAP does not provide formal safety guarantees, the attribution profile can support safety-oriented validation by guiding targeted fault-injection tests. For example, stale, missing, or corrupted boundary frames can be intentionally introduced, followed by verification of the detector response.

Table 9. Average absolute SHAP values in a sample's frames. F(0) is the last frame, F(-4) is the first one

Frame	Avg absolute Shap Values	Diff. vs sample average
$F(0)$	3.34E-04	+5.70%
$F(-1)$	3.00E-04	-5.06%
$F(-2)$	3.03E-04	-4.11%
$F(-3)$	3.13E-04	-0.95%
$F(-4)$	3.31E-04	+4.75%
Avg	3.16E-04	-

4.2. Performance Comparison on Real-World Urban Driving Data with Injected Anomalies

Thus far, the proposed framework has been evaluated on a synthetic dataset, enabling intensive and efficient analysis. To assess its applicability to real-world sensor data, we now present results obtained using the Lyft dataset (Houston et al., 2021), a reference benchmark for LiDAR (Masello et al., 2022), to which artificial noise was added by (Sboui et al., 2025) for the anomaly detection task. Added noise types include simulated multi-path reflections, dynamic synthetic obstacles, uniform weather interference, and abnormal trajectory shifts. Table 10 shows that our proposed encoder-only Transformer (TFMe) (128 embedding size, 5 encoder layers, 8 heads, 2048 FFN and 100 memory slots) outperforms state-of-the-art models, like those presented in (Sboui et al., 2025), stressing the advantage of using the attention mechanism (Vaswani et al., 2017). While the comparison concerns the OOD-aware method (which is the only one employed by the state-of-the-art), we also show that our model employed using the OOD-unaware method still performs better than existing literature models, further highlighting the proposed advancement.

To evaluate the generalization capabilities of the models, we conducted an analysis on an unseen noise type (New noise columns in Table 10). This secondary dataset introduces entirely unseen physical sensor anomalies and hardware faults, such as LiDAR crosstalk (affecting centroid and velocity), GPS Failures (affecting the centroid), sensor occlusion (affecting all features), and sensor saturation (distorting object extent). Upon evaluation with such new noise types, baseline models utilizing OOD-aware thresholds (computed on noise in (Sboui et al., 2025)) exhibited an F1-score degradation between 4% and 6%, while maintaining a constant FPR. On the other hand, the proposed OOD-unaware TFMe Transformer demonstrated superior generalization compared to the baselines and the OOD-aware TFMe Transformer as well, achieving the highest F1-score (93.5%) and the lowest FPR (1.7%). This highlights the quality of the proposed model and the value of the OOD-unaware novelty detection method.

Table 10. Performance comparison of the proposed transformer model with the SoA models on the Lyft Dataset

Model	Approach	Tested on (Sboui et al., 2025)) noise		Tested on new noise	
		F1	FPR	F1	FPR
OC-SVM	OOD-aware (tuned on noise as in (Sboui et al., 2025))	0.476	0.504	0.421	0.504
LOF		0.596	0.440	0.557	0.440
MC-CNN-LSTM AE (Karadayı et al., 2020)		0.815	0.249	0.756	0.249
CNN-BiLSTM VAE (Sboui et al., 2025)		0.826	0.052	0.770	0.052
TFMe Transformer (Ours)		**0.972**	0.043	0.914	0.043
TFMe Transformer (Ours)	OOD-unaware	0.955	**0.017**	**0.935**	**0.017**

Notably, while injection of artificial noise on Lyft makes it possible to investigate a wide variety of noise conditions, we are not aware of any publicly available dataset originally including real-world measured LiDAR anomalies, which would be useful for further progressing the research field.

4.3. Generalization to Unseen Car Park Layouts

The results presented so far were obtained on a single car park layout configuration, and may therefore reflect an implicit specialization of the detectors to the 90° comb-shaped parking layout used for training. To assess this potential limitation, we evaluated the TFMe, ATF-UAD and ATUAD detectors on a 60° angled car park layout. All models were

deployed without retraining and without threshold re-calibration. Table 11 reports the F1-scores at S1 and S5, the lowest and highest severity levels, which respectively represent the least and most detectable operating conditions for each perturbation type, alongside the values obtained on the original layout.

For TFMe, the F1-score decreases from 99% (in the original car park layout) to 98% for ZRB at both severities, from 98-99% to 95-97% for Background, from 88-98% to 86-96% for EMI, and from 88-98% to 85-97% for Occlusion. In no case does the F1-score decrease by more than 3%.

The baseline models are more sensitive to the layout change. ATF-UAD exhibits degradations of 5 percentage points on ZRB and Background at S1 (from 97% to 92% in both cases) and on Occlusion (74% to 69%), and of 8 percentage points on EMI at S1 (from 70% to 62%), the lowest EMI score among all evaluated models. ATUAD maintains competitive performance on ZRB, Background and EMI (reductions between 1% and 6%), but shows the largest degradation of the entire experiment on Occlusion, dropping from 52% to 44% at S1 and from 86% to 73% at S5. The layout change primarily affetct Occlusion for ATUAD and EMI for ATF-UAD, as in the original layout (Section 4.1.2). This suggests that the layout shift seems to widen performance gaps that are already present, rather than introducing new ones.

Table 11. TFMe, ATUAD and ATF-UAD detection performance on an unseen 60° angled layout vs. the original comb-shaped layout

Model	Noise Type	90° Layout		60° Layout	
		S1	S5	S1	S5
TFMe	ZRB	99%	99%	98%	98%
	Background	98%	99%	95%	97%
	EMI	88%	98%	86%	96%
	Occlusion	88%	98%	85%	97%
ATUAD	ZRB	98%	99%	94%	98%
	Background	97%	99%	93%	95%
	EMI	79%	98%	74%	92%
	Occlusion	52%	86%	44%	73%
ATF-UAD	ZRB	97%	99%	92%	97%
	Background	97%	99%	92%	95%
	EMI	70%	93%	62%	87%
	Occlusion	74%	93%	69%	90%

Overall, the results show limited performance degradation under the evaluated layout shift, although they do not establish invariance to arbitrary parking geometries.

4.4. TOR Management

Up to this point, we have analyzed the performance of the models in isolation. However, detectors must be deployed within an ADF vehicular architecture, such as a self-parking system. To assess the impact of the Transformer on the ADF, we conducted targeted experiments incorporating the TOR mechanism, which prompts the driver to resume manual control in the event of potential issues (e.g., (Lee et al., 2020; Soares et al., 2021)).

Accordingly, we expanded the evaluation to include a TOR-management procedure, illustrated in Fig. 2, that suppresses unnecessary activations caused by isolated false positives while repeatedly evaluating the detector output to reduce the risk that persistent anomalies remain undetected. In fact, the window is advanced with a stride of one frame. We implemented a TOR management strategy based on an N-out-of-M logic (Radke and Evanoff, 1994), whereby a TOR is issued only when at least N anomalies are detected within an M-frame sliding window, thus reducing sensitivity to transient noise. To select the proper N, M pair, we performed a false-TOR sensitivity analysis over alternative configurations (N ∈ {1, 2, 3}, M ∈ {5, 8, 10}), as reported in Table 12. The sensitivity analysis was conducted exclusively on ID episodes to avoid using anomaly-domain information for configuration selection. Its objective was to identify the minimum N, M configuration that suppresses unnecessary TORs under nominal operation. Here, N denotes the number of anomalous detections required to trigger a TOR, while M defines the length of the sliding evaluation window. The results indicate that the behavior of the rule is determined mainly by the detection threshold N, whereas the window length M has no observable influence within the examined range. For $N = 1$, a constant ID TOR rate of 1.24% is obtained across all window lengths, since a single false positive is sufficient to trigger a request. Increasing the threshold to N = 2 eliminates all spurious activations, yielding a null ID TOR rate for every value of M. For the ID false-TOR criterion, $N = 3$ provides no additional benefit over $N = 2$, since both produce a zero TOR rate for every tested value of M. Among the tested configurations, (N = 2, M = 5) is the minimum one producing a zero TOR rate on ID episodes, and is adopted for all subsequent experiments.

Table 12. ID false-TOR sensitivity of the N-out-of-M rule (false-TOR rate on ID test episodes)

Minimum Violations [N]	Window Size [M]	ID TOR rate [%]
1	5	1.24
	8	1.24
	10	1.24
2	5	0
	8	0
	10	0
3	5	0
	8	0
	10	0

We then designed an experiment (Figs. 10-13) in which each test case was evaluated across 10,000 randomly selected recorded episodes. The episodes were generated with randomized target parking lots, 0-20 vehicles spawned in the scene, ±5 m lateral offsets, and random 360° initial orientations (Section 3.3). The episodes thus also cover near-boundary normal cases (e.g., minimal clearances arising in dense obstacle configurations) and unusual but valid maneuvers (e.g., large initial misalignments requiring multi-point corrections).

The anomaly detector should remain transparent to ID samples, avoiding unwarranted TORs. Thus, we first analyzed the system under baseline (noise-free) operating conditions and verified that the system achieves a zero TOR rate throughout all test episodes (ID column in the right part of Figs. 10-13, dedicated to the Enhanced ADF), empirically validating the anomaly detector's operational transparency (unimpaired user experience) when the ADF operates within its intended ODD. Then, to simulate varying levels of noise, we injected anomalies into the LiDAR data stream at four different frequencies:

- Scenario A: One frame affected every ten;
- Scenario B: Two consecutive frames affected every ten;
- Scenario C: One frame affected every two;
- Scenario D: Every frame affected.

Under our TOR issue policy, all such scenarios are expected to result in

a TOR, since repeated OOD samples (each including 5 stacked frames) indicate a deviation from the ADF's ODD. Nonetheless, certain noisy inputs may remain within the system's handling capabilities. In such cases, it is acceptable for the maneuver to complete successfully, with the vehicle safely reaching its intended target position. An episode may also terminate due to a time-out, indicating the DRL agent's inability to complete the maneuver. In such scenarios, the ADF initiates a safe, rule-based repositioning maneuver to reset the agent. Alternatively, control may be handed back to the driver, if present.

The left parts of Figs. 10-13 quantify the operational hazards introduced by various noise types in the baseline ADF case (i.e., with no OOD detector nor TOR). Performance should be compared with the noise-free episodes, whose values are represented in the first vertical bar, named ID. The data reveals four distinct impact profiles:

1) ZRB demonstrates the strongest hazard correlation, with success rates declining monotonically as both severity (S1->S5) and injection frequency (Scenario A->D) increase.
2) The impact of Background noise is lower and less dependent on severity level and injection frequency.
3) EMI exhibits its characteristic severity-dependent spectrum (cf. Section 4.1.3), primarily manifesting as timeout proliferation, especially at high severity (Scenarios A & D), rather than direct collisions. This pattern indicates systematic performance degradation rather than immediate safety risks, but with a severe operational efficiency cost.
4) Occlusion generates relatively moderate collision rates, substantially lower than ZRB's impact. This highlights the tendential agreement between detectability and risk (cf. Fig. 5).

In general, it is noticeable that even low-severity levels of noise have an impact on the collision rate. Conversely, our enhanced ADF, introducing the TOR functionality fed by the OOD detector (Fig. 2), prevented collisions in all evaluated episodes and significantly reduced the number of time-outs (Figs. 10-13, right parts). This improvement in safety and robustness comes at the expense of a reduced episode success rate (compensated by issuing TORs). However, it is important to note that the enhanced ADF solution does not issue TORs in the ID case, as can be seen by comparing the ID columns of the right and left parts of Figs. 10-13 (they display the same values).

As discussed in Section 4.1.2, false negatives are non-negligible, substantially because the OOD detector is optimized to minimize FPR. Interestingly, our findings reveal that even when the number of false negatives is large (completion of an episode without TOR means that some noisy samples are not detected), the injected noise does not hinder the agent from successfully completing the parking task (Figs. 10-13, right parts, particularly for Occlusion and low-severity EMI). This suggests that Occlusion and EMI are more difficult to detect, but missed detections do not prevent the agent from reaching the target.

ZRB, Background, and high-severity EMI consistently trigger a TOR across all evaluated scenarios. Occlusion noise is the hardest to detect, even at higher severities, but its impact is limited. This highlights a fundamental trade-off achieved by the enhanced ADF design: the need to trigger a TOR in response to OOD inputs versus the agent's demonstrated ability to achieve its goal handling low-severity OOD scenarios. We argue that an interesting direction to further improve this trade-off would be to extend the ADF's ODD by retraining the system in those OOD conditions that it has shown it can manage effectively.

Fig. 14 analyzes the temporal behavior of the ADF, reporting the average occurrence time of the TOR. At high severity levels, TORs are issued near-instantly (excluding Occlusion cases), while low-severity conditions exhibit significant behavioral variation across noise types. ZRB and Background are quickly detected in all scenarios also at level 1. For the other combinations, system responses look consistent across noise severity levels and scenario frequencies, with Scenario A having the lowest noise frequency and Scenario D the highest. Fig. 14 more effectively illustrates inter-scenario differences than Figs. 10-13, revealing that Scenario A, the sparsest one in terms of noise, diverges from others, particularly regarding the EMI noise.

To analyze the spatial performance, minimum clearance was measured between the ego vehicle and the nearest obstacle, averaged over 10,000 episodes per (noise type, severity) pair (Fig. 15). In the baseline ADF, average minimum clearance decreases with severity level for all perturbation types. Occlusion drops to 0.17 m @S5 against 0.40 m @S1, EMI to 0.32 m against 0.43 m, ZRB to 0.36 m against 0.46 m, and Background to 0.41 m against 0.47 m. Perception degradation therefore narrows the physical safety margin before it produces a collision. This suggests that outcome-level indicators (Figs. 10-13) are not sufficient to estimate the overall operational risk. In the enhanced ADF case, results are opposite to the baseline. Average minimum clearance spans 0.65 m to 1.33 m and increases with severity, because more severe perturbations are detected earlier and the TOR interrupts the maneuver while the vehicle is still distant from surrounding obstacles. This is a valuable behvior given by the system design, particularly (N, M) calibration. The clearance ordering across noise types (ZRB and Background, then EMI, then Occlusion) reflects the detectability ordering of Table 8.

Concluding, these results illustrate the spectrum of potential ADF operating conditions and validate the robustness of the proposed system architecture design across such scenarios.

Fig. 10. ADF performance in the noise scenario A, at different severity levels.

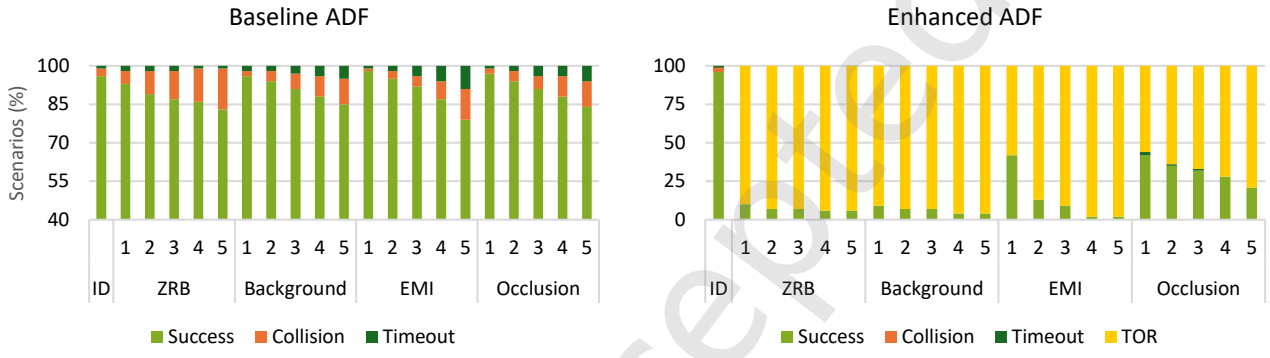

Fig. 11. ADF performance in the noise scenario B, at different severity levels.

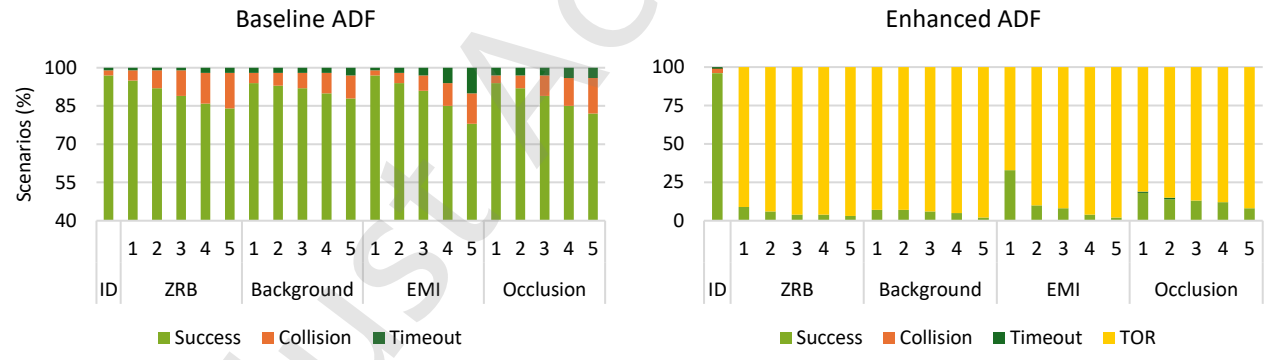

Fig. 12. ADF performance in the noise scenario C, at different severity levels.

Fig. 13. ADF performance in the noise scenario D, at different severity levels.

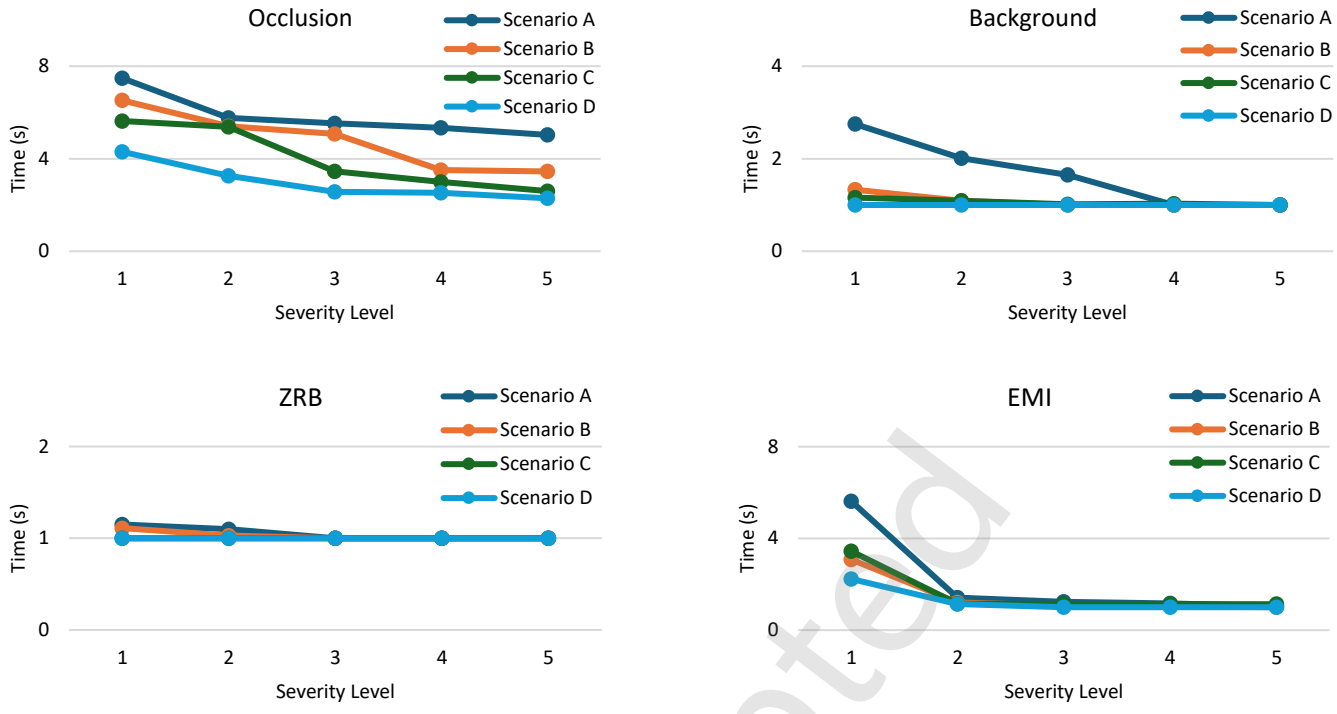

Fig. 14. Average time of First-TOR occurrence (in seconds) by noise type, scenario and severity level.

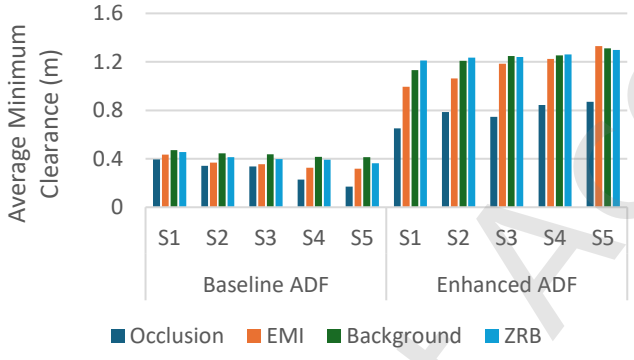

Fig. 15. Average minimum clearance between the ego vehicle and the nearest obstacle by noise type and severity level (S1-S5), for the baseline and enhanced ADF.

4.5. Deployability

Despite being understudied in the literature, deployability analysis remains crucial for AD systems due to their safety-critical embedded nature (Fresta et al., 2025a, 2025b). Performance must therefore be evaluated on resource-constrained platforms (e.g., microcontrollers), with emphasis on latency, power consumption, and energy efficiency (Lv et al., 2023). The decision-making system must operate at 20-50 Hz (Wang et al., 2021) - a constraint that should be kept also after integrating the anomaly detector.

For deployment, we exported the TFMe model into a 1.6 MB INT8 TensorFlow Lite model, while the original DRL agent is deployed as a 10 MB PyTorch model. We adopted TensorFlow Lite for its lightweight runtime efficiency on ARM-based processors. INT8 quantization preserved the performance of the FP32 baseline across all evaluation metrics. The only observed change was for Occlusion at severity level 3, where the F1-score improved from 94% to 95%. Quantization below INT8 was not pursued, as neither the Raspberry Pi 5 nor the Jetson Orin Nano provides hardware acceleration for sub-INT8 data types.

To complement the accuracy comparison of Section 4.1 with the computational profile of the detectors, Table 13 reports, for all compared INT8-quantized models, the multiple-accumulate operations (MACs), operations per inference, size of the exported model file, and memory occupancy per inference. The proposed TFMe requires 4.87M MACs, corresponding to 9.74M operations, a model file size of 1.6 MB and a peak memory usage of 1.34 MB (a footprint that fits the resources of the typical target platforms). This is comparable to ATUAD (4.4M MACs, 8.8M operations, 1.71 MB model file size, 1.68 MB peak memory), indicating that the dual-branch, memory-augmented design introduces only a marginal computational overhead (approximately +10% MACs) while having smaller model size and memory occupancy, and consistently achieving higher detection performance (Section 4.1.2). ATF-UAD is substantially lighter (0.24 M MACs, 0.47 M operations, 0.09 MB model file size, 0.072 MB peak memory), but this efficiency is obtained at the expense of reduced detection accuracy and weaker generalization.

Both models were tested on two embedded boards: NVIDIA Jetson Orin Nano and Raspberry Pi 5. The Jetson Orin Nano integrates a 6-core Arm Cortex-A78AE v8.2 CPU, a 1024-core NVIDIA Ampere architecture GPU with 32 Tensor Cores, and 8 GB LPDDR5 RAM, supporting CUDA for parallelism. The Raspberry Pi 5 features a 2.4 GHz quad-core Cortex-A76 CPU with a VideoCore VII GPU and 8 GB LPDDR4X SDRAM.

On the Jetson Orin Nano, power was monitored via the built-in Jetson power logging utility (*JTOP*), while on the Raspberry Pi 5, it was measured using a USB multimeter connected between the power supply and the board. Energy was computed as the product of average power and inference time, with all results averaged over 1,000 test runs. Thermal impact remains minimal, with both boards showing only a 2-3% increase in temperature. This is essentially due to the 200 ms waiting

Table 13. Computational profile of the compared models

Model	MACs per inference [M]	Operations per inference [M]	Model file size [MB]	Peak memory [MB]
TFMe	4.87	9.74	1.60	1.34
ATUAD	4.40	8.80	1.71	1.68
ATF-UAD	0.24	0.47	0.09	0.07

Table 14. Embedded platforms' performance comparison

Board	Detector				DRL Agent			
	Time [ms]	Power [mW]	Energy [mJ]	Throughput per Watt [IPS/W]	Time [ms]	Power [mW]	Energy [mJ]	Throughput per Watt [IPS/W]
Jetson Orin Nano	1.22	1,450	1.77	565	1.70	1,200	2.04	490
Raspberry Pi 5	1.32	2,500	3.30	303	2.28	950	2.17	462

time between samples, which avoids continuous peak load. Operating temperatures reached roughly 46°C for the Jetson Orin Nano and 62°C for the RP5. Temperatures were measured through the *JTOP* and *vcgencmd* commands, respectively, in a 1-hour period.

Table 14 reports the performance of the two modules running independently on the two embedded platforms. Results show that the detector latency is smaller than thatof the DRL agent. The Jetson Orin Nano reduces the aggregate sequential workload by 19% and the module energy by 30% with respect to the RP5. The gap is given mainly by the Transformer module, where the Raspberry Pi's higher power consumption (2,500 mW) results in an energy usage of 3.30 mJ despite the execution time being comparable. For the self-parking DRL agent, latency is lower on Orin (1.70 vs 2.28 ms), with comparable energy (2.04 vs 2.17 mJ). Moreover, the Jetson Orin Nano has exceptional computational efficiency, achieving a throughput per watt of 565 for the Transformer module, nearly doubling the RP5 (303), while also maintaining a lead in the DRL agent task (490 vs. 462). Table 14 also shows that the enhanced ADF keeps a real-time performance, at the cost of an estimated increase in energy consumption of about 87% on the Jetson Orin Nano and 152% on the Raspberry Pi 5.

5 Conclusion

Since OOD input detection is key for ensuring AI trustworthiness and real-world deployability, we have augmented a self-parking ADF with an OOD detector based on the OOD-unaware novelty detection paradigm. Our advancements over the state-of-the-art, addressing a methodological comparison and a safety-critical innovation gap, include:

- Open-source release of LiDARacks: a toolkit for creating LiDAR distortion models that remain interpretable against sensor parameters and environmental factors https://anonymous.4open.science/r/LiDARacks. Additionally, the source code and model for TFMe are now open source and available at https://anonymous.4open.science/r/TFMe.
- Quantitative sensitivity assessment: evaluated how anomaly detection depends on different noise/attack types and intensities. Results highlight the necessity for robust OOD detection methods with strong generalization properties (OOD-unaware approach, pure novelty detection).
- Better generalization of the OOD-unaware approach: this method provides better generalization capabilities by maintaining consistent performance across various anomaly types and severity levels without the performance drops seen in the OOD-aware case.
- Superiority of the TFMe Transformer: the proposed dual-branch, memory-augmented design outperforms the evaluated baselines.
- Real-world driving data: experiments on authentic on-road Lyft measurements with synthetically injected anomalies show that TFMe performs strongly in both calibration settings and that OOD-unaware calibration transfers more effectively to the evaluated unseen anomaly types.
- Explainability analysis: the temporal attribution profile is distributed across all five frames, with modestly higher average attribution for the first and most recent observations.
- Low intensity noise risks: shown that certain noise types present significant safety hazards during vehicle operation even at low severity levels.
- Effective ADF integration: the anomaly detector improves the performance of a state-of-the-art ADF architecture, mitigating collision risks and timeout occurrences, issuing TORs at repeated anomaly detections. The detector's latency is smaller than that of the self-parking agent.
- Jetson Orin Nano deployment: the board demonstrates greater efficiency than Raspberry Pi 5, achieving lower latency and energy consumption.

We consider these findings instrumental for further advancing research in this area, which is still in its early stages, given the high-stakes nature of automated driving. In our opinion, future work should prioritize both rigorous benchmarking of detection paradigms and fundamental research into unknown anomaly identification. Further research directions may encompass: testing the ADFs beyond self-parking, also featuring a more advanced perception system; extending the evaluation to dynamic agents, alternative sensor configurations, and weather conditions; developing smarter TOR management (e.g., through risk-aware scoring); evaluating more sophisticated attacks (e.g., with temporal correlation) and refining detection models accordingly; assessing Application-Specific Integrated Circuit (ASIC) implementations for optimized embedded performance. The measured tendential agreement between detectability and ADF danger suggests the need for studying risk-aware anomaly scoring, beyond binary detection. Moreover, it would be interesting to study an extension of the ADF's operational design domain to those (currently) OOD conditions that it has shown it can manage effectively, thus

including low-risk noisy samples in the training set to enhance robustness. Finally, real-world deployment should be investigated, to verify the assumptions and refine the models on authentic-context data.

Author contributions

H.B: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **R.B:** Supervision, Methodology, Conceptualization, Funding acquisition, Formal analysis. **L.L:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **A.P:** Software, Data curation. **M.F:** Writing – review & editing, Methodology. **A.S:** Software, Data curation. **V.H:** Formal Analysis, Methodology. **A.D:** Formal Analysis, Methodology. **F.B:** Writing – original draft, Supervision, Methodology, Conceptualization, Funding acquisition, Formal analysis.

Replication and data sharing

The replication package including the code, simulation and data used in this study can be downloaded at https://anonymous.4open.science/r/LiDARacks and https://anonymous.4open.science/r/TFMe.

Acknowledgment

The authors would like to thank all partners within the Hi-Drive project for their cooperation and valuable contribution.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Ao, Z., Su, Y., Li, W., Guo, Q., Zhang, J., 2017. One-Class Classification of Airborne LiDAR Data in Urban Areas Using a Presence and Background Learning Algorithm. Remote Sensing 9, 1001. https://doi.org/10.3390/rs9101001

Ayerdi, J., Iriarte, A., Valle, P., Roman, I., Illarramendi, M., Arrieta, A., 2024. MarMot: Metamorphic Runtime Monitoring of Autonomous Driving Systems. ACM Trans. Softw. Eng. Methodol. 34, 18:1-18:35. https://doi.org/10.1145/3678171

Baccari, S., Hadded, M., Ghazzai, H., Touati, H., Elhadef, M., 2024. Anomaly Detection in Connected and Autonomous Vehicles: A Survey, Analysis, and Research Challenges. IEEE Access 12, 19250–19276. https://doi.org/10.1109/ACCESS.2024.3361829

Bai, X., Dong, P., Wang, J., Huang, Y., Yu, H., Ren, Y., 2025. AdvGLOW: Covert adversarial attacks against autonomous driving perception. Journal of Intelligent and Connected Vehicles 8, 9210067. https://doi.org/10.26599/JICV.2025.9210067

Bellotti, F., Lazzaroni, L., Capello, A., Cossu, M., De Gloria, A., Berta, R., 2023. Explaining a Deep Reinforcement Learning (DRL)-Based Automated Driving Agent in Highway Simulations. IEEE Access 11, 28522–28550. https://doi.org/10.1109/ACCESS.2023.3259544

Berta, R., Lazzaroni, L., Capello, A., Cossu, M., Forneris, L., Pighetti, A., Bellotti, F., 2024. Development of Deep-Learning-Based Autonomous Agents for Low-Speed Maneuvering in Unity. Journal of Intelligent and Connected Vehicles 7, 229–244. https://doi.org/10.26599/JICV.2023.9210039

Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers. SIGMOD Rec. 29, 93–104. https://doi.org/10.1145/335191.335388

Capello, A., Fresta, M., Bellotti, F., Haghighi, H., Hiller, J., Mozaffari, S., Berta, R., 2023. Exploiting Big Data for Experiment Reporting: The Hi-Drive Collaborative Research Project Case. Sensors 23, 7866. https://doi.org/10.3390/s23187866

Chaves-Tibaduiza, J.F., Becerra-Muñoz, A.I., Robledo-Giron, A., Caicedo, O.M., 2024. On the feasibility of using an encoder-only model for anomaly detection: the BERTAD approach, in: 2024 IEEE Colombian Conference on Communications and Computing (COLCOM). Presented at the 2024 IEEE Colombian Conference on Communications and Computing (COLCOM), pp. 1–6. https://doi.org/10.1109/COLCOM62950.2024.10720312

Cheng, C.-H., Huang, C.-H., Brunner, T., Hashemi, V., 2019. Towards Safety Verification of Direct Perception Neural Networks. https://doi.org/10.48550/arXiv.1904.04706

Cheng, Y., Cao, Y., Wang, D., Shen, W., Li, W., 2025. Boosting Global-Local Feature Matching via Anomaly Synthesis for Multi-Class Point Cloud Anomaly Detection. IEEE Transactions on Automation Science and Engineering 1–1. https://doi.org/10.1109/TASE.2025.3544462

Dabbous, A., Berta, R., Fresta, M., Ballout, H., Lazzaroni, L., Bellotti, F., 2024. Bringing Intelligence to the Edge for Structural Health Monitoring: The Case Study of the Z24 Bridge. IEEE Open Journal of the Industrial Electronics Society 5, 781–794. https://doi.org/10.1109/OJIES.2024.3434341

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V., 2017. CARLA: An Open Urban Driving Simulator. https://doi.org/10.48550/arXiv.1711.03938

Fan, J., Wang, Z., Wu, H., Sun, D., Wu, J., Lu, X., 2023. An Adversarial Time–Frequency Reconstruction Network for Unsupervised Anomaly Detection. Neural Networks 168, 44–56. https://doi.org/10.1016/j.neunet.2023.09.018

Fresta, M., Bellotti, F., Bochenko, I., Lazzaroni, L., Merlhiot, G., Tango, F., Berta, R., 2025a. Deep Learning-Based Real-Time Driver Cognitive Distraction Detection. IEEE Access 13, 26589–26607. https://doi.org/10.1109/ACCESS.2025.3539392

Fresta, M., Bellotti, F., Lazzaroni, L., Ballout, H., Pighetti, A., Ameneiro-Prieto, Ó., Porta-Lorenzo, M., Sánchez-Louzán, D., González-Escudeiro, C., Tango, F., Berta, R., 2025b. On-Board Deployability of a Deep Learning-Based System for Distraction and Inattention Detection. IEEE Open Journal of the Industrial Electronics Society 6, 1351–1369. https://doi.org/10.1109/OJIES.2025.3601982

Girdhar, M., Hong, J., Moore, J., 2023. Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models. IEEE Open Journal of Vehicular Technology 4, 417–437. https://doi.org/10.1109/OJVT.2023.3265363

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Van Den Hengel, A., 2019. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised

Anomaly Detection, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1705–1714. https://doi.org/10.1109/ICCV.2019.00179

Gyllenhammar, M., Rodrigues de Campos, G., Törngren, M., 2025. The Road to Safe Automated Driving Systems: A Review of Methods Providing Safety Evidence. IEEE Transactions on Intelligent Transportation Systems 26, 4315–4345. https://doi.org/10.1109/TITS.2025.3532684

He, F., Chen, Y., Chen, R., Nie, W., 2023. Point Cloud Adversarial Perturbation Generation for Adversarial Attacks. IEEE Access 11, 2767–2774. https://doi.org/10.1109/ACCESS.2023.3234313

Henriksen, P., Lomuscio, A., 2020. Efficient Neural Network Verification via Adaptive Refinement and Adversarial Search, in: ECAI 2020. IOS Press, pp. 2513–2520. https://doi.org/10.3233/FAIA200385

Henriksson, J., Berger, C., Ursing, S., Borg, M., 2023. Evaluation of Out-of-Distribution Detection Performance on Autonomous Driving Datasets, in: 2023 IEEE International Conference On Artificial Intelligence Testing (AITest). Presented at the 2023 IEEE International Conference On Artificial Intelligence Testing (AITest), pp. 74–81. https://doi.org/10.1109/AITest58265.2023.00021

Hoppe, I., Hagemann, W., Stierand, I., Hahn, A., Bolles, A., 2024. Challenges for trustworthy autonomous vehicles: Let us learn from life. Systems Engineering 27, 789–800. https://doi.org/10.1002/sys.21744

Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., Omari, S., Iglovikov, V., Ondruska, P., 2021. One Thousand and One Hours: Self-driving Motion Prediction Dataset, in: Kober, J., Ramos, F., Tomlin, C. (Eds.), Proceedings of the 2020 Conference on Robot Learning, Proceedings of Machine Learning Research. PMLR, pp. 409–418.

Javed, A.R., Usman, M., Rehman, S.U., Khan, M.U., Haghighi, M.S., 2021. Anomaly Detection in Automated Vehicles Using Multistage Attention-Based Convolutional Neural Network. IEEE Transactions on Intelligent Transportation Systems 22, 4291–4300. https://doi.org/10.1109/TITS.2020.3025875

Jin, Z., Jiang, Q., Lu, X., Yan, C., Ji, X., Xu, W., 2024. PhantomLiDAR: Cross-modality Signal Injection Attacks against LiDAR. https://doi.org/10.14722/ndss.2025.23997

Kacianka, S., Pretschner, A., 2021. Designing Accountable Systems, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21. Association for Computing Machinery, New York, NY, USA, pp. 424–437. https://doi.org/10.1145/3442188.3445905

Karadayı, Y., Aydin, M.N., Öğrenci, A.S., 2020. A Hybrid Deep Learning Framework for Unsupervised Anomaly Detection in Multivariate Spatio-Temporal Data. Applied Sciences 10. https://doi.org/10.3390/app10155191

Kingma, D.P., Welling, M., 2022. Auto-Encoding Variational Bayes. https://doi.org/10.48550/arXiv.1312.6114

Kosuge, A., Yu, L., Hamada, M., Matsuo, K., Kuroda, T., 2023. A Deep Metric Learning-Based Anomaly Detection System for Transparent Objects Using Polarized-Image Fusion. IEEE Open Journal of the Industrial Electronics Society 4, 205–213. https://doi.org/10.1109/OJIES.2023.3284014

Kumar, B., Sinha, A., Chakrabarti, S., Vyas, O.P., 2021. A fast learning algorithm for One-Class Slab Support Vector Machines. Knowledge-Based Systems 228, 107267. https://doi.org/10.1016/j.knosys.2021.107267

Lazzaroni, L., Pighetti, A., Bellotti, F., Capello, A., Cossu, M., Berta, R., 2024. Automated Parking in CARLA: A Deep Reinforcement Learning-Based Approach, in: Bellotti, F., Grammatikakis, M.D., Mansour, A., Ruo Roch, M., Seepold, R., Solanas, A., Berta, R. (Eds.), Applications in Electronics Pervading Industry, Environment and Society. Springer Nature Switzerland, Cham, pp. 352–357. https://doi.org/10.1007/978-3-031-48121-5_50

Lee, C.W., Nayeer, N., Garcia, D.E., Agrawal, A., Liu, B., 2020. Identifying the Operational Design Domain for an Automated Driving System through Assessed Risk, in: 2020 IEEE Intelligent Vehicles Symposium (IV). Presented at the 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 1317–1322. https://doi.org/10.1109/IV47402.2020.9304552

Li, H., Zhang, C., Lin, L., Guo, L., Li, R., Zhao, H., Di, L., 2024. In-Season Mapping of Sugarcane Planting Based on Sentinel-2 Imagery. https://doi.org/10.2139/ssrn.4808169

Li, Y., Wen, C., Juefei-Xu, F., Feng, C., 2021. Fooling LiDAR Perception via Adversarial Trajectory Perturbation, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp. 7878–7887. https://doi.org/10.1109/ICCV48922.2021.00780

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation Forest, in: 2008 Eighth IEEE International Conference on Data Mining. Presented at the 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. https://doi.org/10.1109/ICDM.2008.17

Liu, Z., Lin, F., Meng, T., Baha-eddine, B.C., Lu, L., Xue, Q., Ren, K., 2024. EMTrig: Physical Adversarial Examples Triggered by Electromagnetic Injection towards LiDAR Perception, in: Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems, SenSys '24. Association for Computing Machinery, New York, NY, USA, pp. 351–364. https://doi.org/10.1145/3666025.3699343

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp. 4768–4777.

Luo, W., Yao, H., Yu, W., Li, Z., 2025. AMI-Net: Adaptive Mask Inpainting Network for Industrial Anomaly Detection and Localization. IEEE Transactions on Automation Science and Engineering 22, 1591–1605. https://doi.org/10.1109/TASE.2024.3368142

Lv, W., Yang, P., Zheng, T., Yi, B., Ding, Y., Wang, Q., Deng, M., 2023. Energy Consumption and QoS-Aware Co-Offloading for Vehicular Edge Computing. IEEE Internet of Things Journal 10, 5214–5225. https://doi.org/10.1109/JIOT.2022.3221966

Masello, L., Sheehan, B., Murphy, F., Castignani, G., McDonnell, K., Ryan, C., 2022. From Traditional to Autonomous Vehicles: A Systematic Review of Data Availability. Transportation Research

Record 2676, 161–193. https://doi.org/10.1177/03611981211057532
Mavikumbure, H.S., Cobilean, V., Wickramasinghe, C.S., Varghese, B.J., Carlson, R.B., Rieger, C., Pennington, T., Manic, M., 2024. Cy-Phy ADS: Cyber-Physical Anomaly Detection Framework for EV Charging Systems. IEEE Transactions on Transportation Electrification 10, 9904–9917. https://doi.org/10.1109/TTE.2024.3363672
Pham, T.M.T., Yang, B., Yang, J., 2024. Evaluating the Robustness of LiDAR-based 3D Obstacles Detection and Its Impacts on Autonomous Driving Systems. https://doi.org/10.48550/arXiv.2408.13653
Prathiba, S.B., Raja, G., Anbalagan, S., S, A.K., Gurumoorthy, S., Dev, K., 2023. A Hybrid Deep Sensor Anomaly Detection for Autonomous Vehicles in 6G-V2X Environment. IEEE Transactions on Network Science and Engineering 10, 1246–1255. https://doi.org/10.1109/TNSE.2022.3188304
Qiao, Y., Wu, K., Jin, P., 2023. Efficient Anomaly Detection for High-Dimensional Sensing Data With One-Class Support Vector Machine. IEEE Transactions on Knowledge and Data Engineering 35, 404–417. https://doi.org/10.1109/TKDE.2021.3077046
Radke, G.E., Evanoff, J., 1994. A fast recursive algorithm to compute the probability of M-out-of-N events, in: Proceedings of Annual Reliability and Maintainability Symposium (RAMS). Presented at the Annual Reliability and Maintainability Symposium (RAMS), pp. 114–117. https://doi.org/10.1109/RAMS.1994.291092
Rodríguez-Arozamena, M., Matute, J., Araluce, J., Kuschnig, L., Pilz, C., Schratter, M., Rastelli, J.P., Zubizarreta, A., 2025. A Fallback Localization Algorithm for Automated Vehicles Based on Object Detection and Tracking. IEEE Open Journal of Vehicular Technology 6, 1085–1099. https://doi.org/10.1109/OJVT.2025.3560198
Rowe, F., Jeanneret Medina, M., Journé, B., Coëtard, E., Myers, M., 2024. Understanding responsibility under uncertainty: A critical and scoping review of autonomous driving systems. Journal of Information Technology 39, 587–615. https://doi.org/10.1177/02683962231207108
SAE International, 2014. J3016_201401 : Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems [WWW Document]. URL https://www.sae.org/standards/j3016_201401-taxonomy-definitions-terms-related-road-motor-vehicle-automated-driving-systems (accessed 4.15.26).
Sboui, N., Ghazzai, H., Hadded, M., Elhadef, M., Setti, G., 2025. Unsupervised Hybrid VAE-Based Anomaly Detection for Vehicle Onboard LiDAR Sensors. IEEE Access 13, 81835–81847. https://doi.org/10.1109/ACCESS.2025.3568395
Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X., 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. ACM Trans. Database Syst. 42, 19:1-19:21. https://doi.org/10.1145/3068335
Shapley, L.S., 1952. A Value for N-Person Games. RAND Corporation.
Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H., 2020. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp. 10526–10535. https://doi.org/10.1109/CVPR42600.2020.01054
Soares, S., Lobo, A., Ferreira, S., Cunha, L., Couto, A., 2021. Takeover performance evaluation using driving simulation: a systematic review and meta-analysis. European Transport Research Review 13, 47. https://doi.org/10.1186/s12544-021-00505-2
Sun, J., Zhang, Q., Kailkhura, B., Yu, Z., Xiao, C., Mao, Z.M., 2022. Benchmarking Robustness of 3D Point Cloud Recognition Against Common Corruptions. https://doi.org/10.48550/arXiv.2201.12296
Swana, E.F., Bokoro, P., Doorsamy, W., 2024. Isolation Forest and Local Outlier Factor-Based Anomaly Detection on an Induction Machine, in: 2024 IEEE PES/IAS PowerAfrica. Presented at the 2024 IEEE PES/IAS PowerAfrica, pp. 1–5. https://doi.org/10.1109/PowerAfrica61624.2024.10759503
Trivedi, C., Bhattacharya, P., Prasad, V.K., Patel, V., Singh, A., Tanwar, S., Sharma, R., Aluvala, S., Pau, G., Sharma, G., 2024. Explainable AI for Industry 5.0: Vision, Architecture, and Potential Directions. IEEE Open Journal of Industry Applications 5, 177–208. https://doi.org/10.1109/OJIA.2024.3399057
van Wyk, F., Wang, Y., Khojandi, A., Masoud, N., 2020. Real-Time Sensor Anomaly Detection and Identification in Automated Vehicles. IEEE Transactions on Intelligent Transportation Systems 21, 1264–1276. https://doi.org/10.1109/TITS.2019.2906038
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. ukasz, Polosukhin, I., 2017. Attention is All you Need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.
Wang, J., Li, F., Zhang, X., Sun, H., 2024. Adversarial Obstacle Generation Against LiDAR-Based 3D Object Detection. IEEE Transactions on Multimedia 26, 2686–2699. https://doi.org/10.1109/TMM.2023.3302018
Wang, X., Qi, X., Wang, P., Yang, J., 2021. Decision making framework for autonomous vehicles driving behavior in complex scenarios via hierarchical state machine. Auton. Intell. Syst. 1, 10. https://doi.org/10.1007/s43684-021-00015-x
Xu, H., Zhang, L., Li, P., Zhu, F., 2022. Outlier detection algorithm based on k-nearest neighbors-local outlier factor. Journal of Algorithms & Computational Technology 16, 17483026221078111. https://doi.org/10.1177/17483026221078111
Yang, D., Cai, X., Liu, Z., Jiang, W., Zhang, B., Yan, G., Gao, X., Liu, S., Shi, B., 2024. Realistic Rainy Weather Simulation for LiDARs in CARLA Simulator, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Presented at the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 951–957. https://doi.org/10.1109/IROS58592.2024.10802036
Yang, Y., Wang, J., Guo, X., Yang, X., Qin, W., 2025. Methods for Improving Point Cloud Authenticity in LiDAR Simulation for Autonomous Driving: A Review. IEEE Access 13, 4562–4580. https://doi.org/10.1109/ACCESS.2025.3525805
Yu, D., Deng, L., 2015. Gaussian Mixture Models, in: Yu, D., Deng, L. (Eds.), Automatic Speech Recognition: A Deep Learning Approach. Springer, London, pp. 13–21.

https://doi.org/10.1007/978-1-4471-5779-3_2

Yu, X., Zhang, K., Liu, Y., Zou, B., Wang, J., Wang, W., Qian, R., 2025. Adversarial Transformer-Based Anomaly Detection for Multivariate Time Series. IEEE Transactions on Industrial Informatics 21, 2471–2480. https://doi.org/10.1109/TII.2024.3507211

Yu, Z., Yang, T., Chang, Q., Liu, Y., Wang, W., 2024. Attribution-Based Scanline Perturbation Attack on 3d Detectors of Lidar Point Clouds, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4570–4574. https://doi.org/10.1109/ICASSP48485.2024.10447340

Zhang, C., Han, J., Zou, Y., Dong, K., Li, Y., Ding, J., Han, X., 2024. Detecting the Anomalies in LiDAR Pointcloud. pp. 2024-01–2045. https://doi.org/10.4271/2024-01-2045

Zhang, Y., Tu, C., Gao, K., Wang, L., 2024. Multisensor Information Fusion: Future of Environmental Perception in Intelligent Vehicles. Journal of Intelligent and Connected Vehicles 7, 163–176. https://doi.org/10.26599/JICV.2023.9210049

Zhao, J., Li, Y., Zhu, B., Deng, W., Sun, B., 2021. Method and Applications of Lidar Modeling for Virtual Testing of Intelligent Vehicles. IEEE Transactions on Intelligent Transportation Systems 22, 2990–3000. https://doi.org/10.1109/TITS.2020.2978438

Zhirong Wu, Song, S., Khosla, A., Fisher Yu, Linguang Zhang, Xiaoou Tang, Xiao, J., 2015. 3D ShapeNets: A deep representation for volumetric shapes, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, pp. 1912–1920. https://doi.org/10.1109/CVPR.2015.7298801

Zhou, Q., Feng, Z., Gu, Q., Cheng, G., Lu, X., Shi, J., Ma, L., 2022. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. Computer Vision and Image Understanding 221, 103448. https://doi.org/10.1016/j.cviu.2022.103448

Zhu, L., Borlaug, B., Lin, L., Holden, J., Gonder, J., 2021. Identifying Light-Duty Vehicle Travel from Large-Scale Multimodal Wearable GPS Data with Novelty Detection Algorithms, in: 2021 IEEE Green Technologies Conference (GreenTech). Presented at the 2021 IEEE Green Technologies Conference (GreenTech), pp. 155–162. https://doi.org/10.1109/GreenTech48523.2021.00034

Hadi Ballout is a Ph.D. student at ELIOS Lab, DITEN, University of Genoa, Genoa, Italy. He obtained the B.Sc. degree in Computer Engineering and the M.Sc. degree in Computer and Communication Engineering from Lebanese International University, Lebanon, in 2019 and 2021 respectively. His research interests include embedded machine and deep learning, explainable AI, and automotive.

Riccardo Berta is Associate Professor at ELIOS Lab (DITEN), University of Genoa. His main research interest is the application of electronic systems, in the fields of Machine Learning, Internet of Things (IoT) and Serious Gaming. He is founding member of the Serious Games Society. He authored 170+ papers in international journals and conferences. He is Publications Chair for the Applications in Electronics Pervading Industry, Environment and Society (ApplePies) International Conference.

Luca Lazzaroni received the B.Sc. degree in electronic engineering and information technologies, the M.Sc. degree in electronic engineering, and the Ph.D. degree in science and technology for electronic and telecommunication engineering (STIET) from the University of Genoa, Genoa, Italy, in 2017, 2020, and 2024, respectively. He is currently an Assistant Professor with the ELIOS Laboratory, Department of Electrical, Electronic, Telecommunication Engineering and Naval Architecture, University of Genoa. He has co-authored about 50 papers in international journals and conferences. His research interests include deep learning, tiny machine learning, and automated driving.

Alessandro Pighetti is a Ph.D. student at ELIOS Laboratory, University of Genoa, currently enrolled in a Joint Doctoral Programme in Interactive and Cognitive

Environments (JD-ICE) in collaboration with the University of Genoa (UniGe) and the Queen Mary University of London (QMUL). He earned his B.Sc. (2019) and M.Sc. (2022) in Computer Science and Data Science & Engineering: Artificial Intelligence respectively from the University of Genoa. His research includes deep learning, tiny machine learning, serious games, and automated driving.

Matteo Fresta is a Ph.D. student at ELIOS Lab, DITEN, University of Genoa, Genoa, Italy. He obtained the B.Sc. degree in Electronic Engineering and Information Technologies in 2019 and the M.Sc. degree in Electronic Engineering in 2022 from the University of Genoa. His research activities include deep learning, IoT, tiny machine learning and automotive.

Ammar Saad is a Ph.D. student at ELIOS Lab, DITEN, University of Genoa, Genoa, Italy. He obtained his B.Sc. and M.Sc. degrees in Electronics Engineering from Lebanese International University, Lebanon, in 2019 and 2021 respectively. His research interests include accelerating machine and deep learning models, generative AI, and FPGA.

Vahid Hashemi holds a Ph.D. (Dr.-Ing.) in Computer Science from the Max Planck Institute for Informatics and Saarland University. He currently leads the Product Data & AI Program within the Software Platform and Integration division at AUDI AG. Prior to this role, he served as a Technical Team Lead, where he was responsible for developing advanced methods and tools for the safety verification and validation of autonomous driving systems. Dr. Hashemi has served as a technical lead in several federally funded research projects, including EU Hi-Drive and KARLI, and has contributed as an advisory board member to the EU RobustifAI and FOCETA projects. He is also actively involved in international standardization and policy initiatives: he is a committee member of the SAE G34 working group on system and safety considerations for machine learning, and a member of the OECD.AI Expert Group on AI Risk & Accountability Management.

Akshay Dhonthi Ramesh Babu received the M.Sc. degree (cum laude) in artificial intelligence and robotics from Sapienza university of Rome, and the Ph.D. degree in safety assurance of autonomous systems from the University of Twente, in 2022 and 2025, respectively. His research lies at the intersection of artificial intelligence, formal methods, and robotics, focusing on ensuring the safety and robustness of AI-driven systems. His doctoral studies at the University of Twente were in collaboration with AUDI AG since 2022, where his research focused on the safety assurance of AI systems, with applications in autonomous vehicles. During his PhD, Akshay contributed to several international conferences, with notable work on neural network security, temporal logic, and reinforcement learning verification.

Francesco Bellotti received the M.Sc. degree (cum laude) in electronic engineering and the Ph.D. degree in electrical engineering from the University of Genoa, in 1997 and 2001, respectively. He is an Associate Professor with DITEN, University of Genoa, teaching cyber-physical systems, computational intelligence and machine learning for automated driving. He has co-authored 280+ scientific publications in journals, international books, and conference proceedings. He has been responsible for the design

and implementation WPs of several industrial research projects, particularly in big data and automotive. His main research interests include automated driving, machine learning, embedded systems.