remote sensing

MDPI

Article

EAS-Det: Edge-Aware Semantic Feature Fusion for Robust 3D Object Detection in LiDAR Point Clouds

Huishan Wang [1], Jie Ma [1,*], Yuehua Zhao [2], Jianlei Zhang [1] and Fangwei Chen [1]

[1] School of Electronics and Information Engineering, Hebei University of Technology, Tianjin 300401, China; 202211901015@stu.hebut.edu.cn (H.W.); 202321902031@stu.hebut.edu.cn (J.Z.); 202331905073@stu.hebut.edu.cn (F.C.)

[2] College of Electrical Engineering, North China University of Science and Technology, Tangshan 063210, China; yuehua@ncst.edu.cn

* Correspondence: jma@hebut.edu.cn

Highlights

What are the main findings?

- A novel edge-aware semantic fusion framework (EAS-Det) is proposed to enhance LiDAR-based 3D object detection via multi-scale geometric–semantic interaction.
- The dual-attention ESI module adaptively fuses edge and semantic cues, improving boundary precision and detection performance for small objects such as pedestrians and cyclists.

What are the implications of the main findings?

- EAS-Det achieves significant AP gains on the KITTI and Waymo benchmarks while maintaining real-time efficiency.
- The lightweight and modular design of EAS-Det allows for effortless integration into mainstream 3D detection backbones for real-world applications.

Abstract

Accurate 3D object detection and localization in LiDAR point clouds are crucial for applications such as autonomous driving and UAV-based monitoring. However, existing detectors often suffer from the loss of critical geometric information during network processing, mainly due to downsampling and pooling operations. This leads to imprecise object boundaries and degraded detection accuracy, particularly for small objects. To address these challenges, we propose Edge-Aware Semantic Feature Fusion for Detection (EAS-Det), a lightweight, plug-and-play framework for LiDAR-based perception. The core module, Edge-Semantic Interaction (ESI), employs a dual-attention mechanism to adaptively fuse geometric edge cues with high-level semantic context, yielding multi-scale representations that preserve structural details while enhancing contextual awareness. EAS-Det is compatible with mainstream backbones such as PointPillars and PV-RCNN. Extensive experiments on the KITTI and Waymo datasets demonstrate consistent and significant improvements, achieving up to 10.34% and 8.66% AP gains for pedestrians and cyclists, respectively, on the KITTI benchmark. These results underscore the effectiveness and generalizability of EAS-Det for robust 3D object detection in complex real-world environments.

Keywords: point clouds; feature fusion; edge-semantic interaction; object detection

check for updates

Academic Editors: Wensheng Wang, Martin Gade, Yi Wan and Zhirui Wang

Received: 20 October 2025
Revised: 13 November 2025
Accepted: 15 November 2025
Published: 18 November 2025

Citation: Wang, H.; Ma, J.; Zhao, Y.; Zhang, J.; Chen, F. EAS-Det: Edge-Aware Semantic Feature Fusion for Robust 3D Object Detection in LiDAR Point Clouds. *Remote Sens.* **2025**, *17*, 3743. https://doi.org/10.3390/rs17223743

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The widespread adoption of LiDAR and multi-sensor platforms has established 3D point clouds as a cornerstone for environmental perception in applications such as autonomous driving. While UAV-based monitoring represents a promising future extension of this research, the present study focuses on ground-level LiDAR perception for urban road scenes. Among various perception tasks, 3D object detection and localization are particularly crucial, as they provide essential spatial information for downstream applications such as path planning, dynamic monitoring, and disaster response. Reliable object identification and localization in complex environments form the foundation for dynamic scene interpretation, supporting intelligent systems in autonomous navigation and cooperative sensing. Moreover, accurate perception of small, occluded, or non-cooperative objects is critical for ensuring operational safety and facilitating effective data fusion across sensing platforms.

Despite remarkable progress in deep learning-based 3D detection [1–3], several challenges persist under real-world conditions. Severe occlusion, edge ambiguity, and the detection of small or non-cooperative objects in cluttered scenes remain major obstacles. Pioneering work such as PointNet [4] introduced direct point-based feature learning, while voxel-based methods like VoxelNet [5] and SECOND [6] enhanced spatial representation through 3D convolution on structured grids. Subsequent multi-view frameworks, including MV3D [7] and PointPillars [8], aimed to achieve a better balance between accuracy and efficiency by projecting 3D data into 2D spaces. Meanwhile, multi-stage detectors such as PointRCNN [9] and RSN [10] adopted coarse-to-fine refinement pipelines [11]. Nevertheless, a common limitation of these approaches is the progressive loss of fine geometric details due to downsampling and pooling operations. This degradation in localization precision is particularly pronounced along object boundaries and for small or occluded instances, as illustrated in Figure 1. These limitations not only reduce detection accuracy but also hinder downstream tasks such as object tracking, motion prediction, and multi-source data fusion. Although evaluated on autonomous driving datasets such as KITTI [12] and Waymo [13], we emphasize that these benchmarks represent ground-based LiDAR sensing environments that share fundamental challenges with remote sensing, including sparse sampling, occlusion, and background clutter. Consequently, the advancements achieved in this demanding near-range domain underscore the potential of our method for broader remote sensing applications, particularly for large-scale 3D scene reconstruction and spatial structure understanding across diverse viewing geometries.

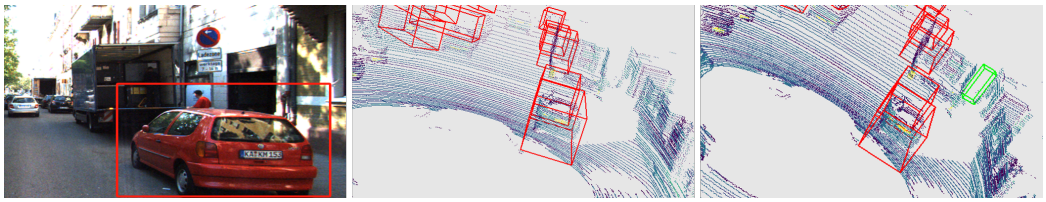

Figure 1. A complex scenario where a car (red bounding box) and a pedestrian (green bounding box) overlap (**left**). Conventional methods often fail under background clutter (**center**), whereas EAS-Det achieves accurate detection and precise localization even under occlusion (**right**).

To overcome these limitations, we propose Edge-Aware Semantic Feature Fusion for Detection (EAS-Det), a lightweight and plug-and-play framework that explicitly integrates geometric edge cues with high-level semantic features to enhance structural awareness and contextual reasoning; it can be seamlessly integrated with various 3D detection backbones without modifying their internal architectures. Unlike existing approaches that simply concatenate semantic and geometric features, EAS-Det introduces a bidirectional

edge–semantic attention mechanism within the ESI module. This design adaptively reweights low-level geometric gradients and high-level semantic cues, achieving dynamic feature balancing across multiple receptive fields. At the core of our framework lies the Edge–Semantic Interaction (ESI) module, which employs a dual-attention strategy to produce structure-preserving and context-aware representations. By effectively capturing fine-grained geometric details together with semantic context, the framework exhibits strong robustness, achieving substantial performance gains in the detection of small objects like pedestrians and cyclists, which are particularly challenging in sparse point clouds.

The ESI module comprises three complementary branches: (1) a geometric branch, which projects the raw point cloud into the bird's-eye view (BEV) and applies Sobel filtering to extract pixel-wise gradient information, thereby capturing salient edge structures; (2) a fusion branch, which employs a channel–spatial attention mechanism to adaptively weight and integrate raw and edge-enhanced features, yielding context-rich multi-scale representations; and (3) a semantic branch, which processes edge-augmented features through a 3D semantic encoder to produce high-level semantic maps with confidence scores. The outputs from these branches are combined into a unified representation that improves detection accuracy and robustness while providing a reliable foundation for temporal consistency in multi-frame tracking and cross-platform perception.

We integrate EAS-Det into several mainstream 3D detection backbones, including PointPillars, PointRCNN, and PV-RCNN [14], and evaluate it on the KITTI and Waymo benchmarks. Experimental results show consistent performance improvements, particularly for small and heavily occluded objects, confirming that EAS-Det effectively preserves geometric integrity and semantic coherence.

In summary, the main contributions of this work are as follows:

- We propose EAS-Det, a lightweight and plug-and-play framework for fine-grained 3D object detection, which integrates multi-scale edge, geometric, and semantic features to enhance spatial perception and localization accuracy.
- We design an ESI module that employs a dual-attention mechanism to adaptively integrate edge and semantic cues, improving boundary precision, contextual understanding, and temporal consistency in complex and cluttered environments.
- Comprehensive experiments on the KITTI and Waymo datasets demonstrate the superior accuracy, robustness, and generalization of EAS-Det, particularly for small objects.

2. Related Work

Deep learning has significantly advanced the field of 3D object detection. Existing methods can be broadly categorized into two major paradigms: two-stage detectors and one-stage detectors. These paradigms embody distinct design philosophies, aiming to balance accuracy, efficiency, and robustness under diverse sensing conditions.

Two-stage detectors typically generate candidate proposals before refining classification and localization. In 2D vision, pioneering methods like R-CNN [15], Fast R-CNN [16], Faster R-CNN [17], and Mask R-CNN [18] established the proposal-based paradigm, which enables high detection accuracy through iterative region refinement. Extending this to 3D perception, PV-RCNN [14] effectively fuses voxel and point features using a novel voxel set abstraction module, while CT3D [19] employs transformer-based encoders to capture long-range dependencies and spatial correlations. Building on these principles, methods like AVOD-FPN [20] and DetectoRS [21] emphasize multi-scale fusion and small-object detection via hierarchical feature pyramid integration. Recent advances in sparse representation, such as VoxelNeXt [22] and SparseBEV [23], have introduced fully sparse 3D convolutional networks, significantly improving detection efficiency while maintaining accuracy. Although these methods achieve state-of-the-art performance, they involve sub-

stantial computational costs and are prone to error propagation from inaccurate region proposals. This limits their deployment on real-time or resource-limited remote sensing platforms such as UAVs and satellites, highlighting the need for approaches that maintain accuracy while reducing computational complexity.

One-stage detectors, in contrast, perform classification and regression in a single pass, offering faster inference suitable for real-time applications. Representative 2D frameworks include YOLO [24–26] and SSD [27], while for 3D point clouds, methods like VoxelNet and SECOND employ sparse 3D convolutional networks on voxelized representations. PointPillars encodes point clouds into pseudo-images for efficient 2D convolution, and CenterPoint [28] detects objects directly from heatmap-predicted centroids in an anchor-free manner. More recent approaches, such as IA-SSD [29], introduce instance-aware downsampling to preserve critical points, and FAR-Pillar [30] employs feature adaptive refinement for oriented 3D detection. These approaches strike a balance between efficiency and accuracy but often lack fine-grained spatial granularity and explicit boundary modeling. Consequently, their performance suffers on small, occluded, or edge-ambiguous objects, which are prevalent in remote sensing imagery where object scale, density, and viewpoint vary drastically. The absence of explicit geometric modeling in these methods highlights the importance of incorporating structural priors for robust detection in complex environments.

Fusion-based methods further enhance perception by integrating complementary modalities at various stages of the detection pipeline. Early fusion approaches like F-PointNet [31] leverage 2D detections to constrain the 3D search space through frustum estimation, while PointPainting [32] augments LiDAR points with semantic predictions from RGB images in a sequential pipeline. More recently, SeSame [33] learns semantic features directly from point clouds through a dedicated segmentation head, reducing dependency on auxiliary modalities. Advanced fusion strategies such as VoxelNextFusion [34] implement unified voxel-based cross-modal integration, and Fast-CLOCs [35] demonstrate improved real-time performance through efficient camera-LiDAR candidate fusion. Despite these advances, fusion-based strategies remain sensitive to the quality, temporal alignment, and calibration accuracy of multimodal data, which limits their scalability in heterogeneous systems like satellite–aerial–ground sensor networks. Moreover, reliance on auxiliary sensors or annotations reduces generalization to resource-limited or single-modality settings, highlighting the value of approaches that achieve robust performance using LiDAR data alone.

In summary, two-stage detectors achieve high accuracy at substantial computational costs, one-stage methods offer real-time efficiency with limited robustness, and fusion-based frameworks enhance semantic understanding while introducing external dependencies. These limitations highlight the need for a unified approach that integrates geometric fidelity and semantic context from point cloud data. To address this gap, the proposed EAS-Det framework incorporates an ESI module that explicitly integrates geometric boundary cues with high-level semantic features through a dual-attention mechanism. By preserving structural details while enhancing discriminative power through adaptive feature re-weighting, EAS-Det achieves accurate and robust 3D detection suitable for large-scale, heterogeneous remote sensing environments.

3. Method Overview

The proposed EAS-Det architecture processes raw LiDAR point clouds to predict oriented 3D bounding boxes, effectively integrating geometric, edge, and semantic cues for robust detection in complex environments. As illustrated in Figure 2, the framework comprises three main stages: (1) Feature Extraction and Interaction through the ESI module, which extracts and balances high-precision edge and semantic features using a

dual-attention mechanism; (2) Feature Concatenation and Fusion, where multi-scale representations are constructed by aligning and aggregating geometric, edge, and semantic cues; (3) LiDAR-based 3D Object Detection, where fused features are fed into a detection head for oriented 3D bounding box prediction, enabling accurate localization and classification under challenging conditions.

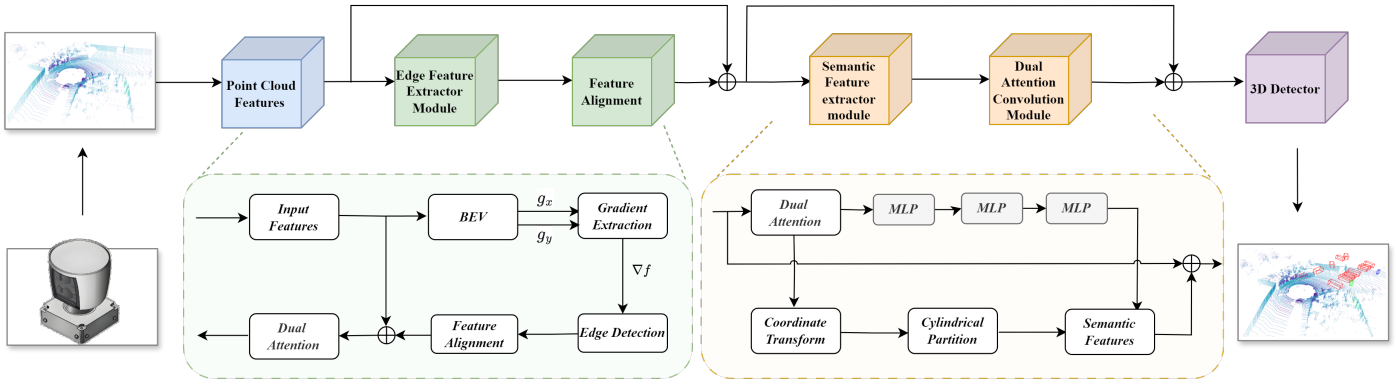

Figure 2. Overview of the EAS-Det framework. Raw point clouds are processed through the Edge-Semantic Interaction (ESI) module, fused, and input to a LiDAR-based detection head for oriented 3D bounding box prediction.

3.1. Feature Extraction and Interaction

Robust 3D object detection requires capturing both fine-grained geometric structures and high-level semantic context. Relying solely on geometric cues or semantic information proves inadequate, particularly in remote sensing scenarios characterized by sparse, occluded, or cluttered point clouds. To address this challenge, the proposed ESI module integrates three interdependent components: the Edge Feature Extraction (EF) Module, the Dual Attention (DA) Module, and the 3D Semantic Feature Extraction (SF) Module. By explicitly modeling geometric edges and semantic representations through attention-driven interaction, EAS-Det effectively captures both local object boundaries and global contextual dependencies.

3.1.1. Edge Feature Extraction Module

The EF module extracts precise geometric edge features encoding local structure and object contours, which are critical for detecting small, thin, or sparsely represented objects. The raw point cloud is denoted as:

$$X = \{x_i \mid 1 \leq i \leq N\}, \tag{1}$$

and is initially downsampled to N_1 points using farthest point sampling (FPS) to reduce computational cost while preserving spatial coverage. Each point is projected onto a bird's-eye view (BEV) grid, where cells aggregate statistics including maximum height, intensity, and point density. This BEV representation facilitates efficient 2D convolution operations while maintaining spatial locality.

Edge detection employs Sobel operators along the x and y directions:

$$g_x = f(x,y) * k_x, \quad g_y = f(x,y) * k_y, \tag{2}$$

where $*$ denotes convolution, with convolution kernels:

$$k_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad k_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \tag{3}$$

Gradient magnitude and orientation are computed as:

$$M(x,y) \approx |g_x| + |g_y|, \quad \alpha(x,y) = \arctan\left(\frac{g_y}{g_x}\right). \tag{4}$$

Non-maximum suppression (NMS) and dual-thresholding refine the edge maps. To accommodate objects at varying scales, multi-scale BEV grids generate edge maps denoted as $\{M^{(s)}(x,y) \mid s \in \{1,2,3\}\}$, which are fused to provide scale-invariant edge cues.

Through lightweight operations including FPS sampling, BEV projection, and 3×3 Sobel convolution, the EF module achieves robust feature extraction with minimal overhead, enabling real-time deployment.

While the Sobel operator originates from image gradient computation, in this context the term edge specifically denotes two-dimensional gradient features computed on the BEV height map that correspond to object contours or elevation discontinuities in the 3D scene. This approach enables the extracted edge cues to serve as a compact yet physically meaningful approximation of local surface variations. Moreover, this gradient-based extraction exhibits inherent robustness to point cloud sparsity, as it relies on relative intensity contrasts between adjacent voxels rather than absolute point density. This property enables the preservation of salient edge structures even in sparse BEV projections, which is crucial for maintaining detection performance under challenging sensing conditions.

3.1.2. Dual Attention Convolution Module

The Dual Attention (DA) module enhances feature representations by selectively emphasizing informative channels and spatial regions, enabling precise boundary delineation and robust detection of small or sparse objects. This module integrates edge and semantic features from preceding layers, capturing both geometric and contextual information essential for 3D object detection.

The fused feature tensor obtained from initial edge and point feature mapping is denoted as:

$$F_n \in \mathbb{R}^{N \times C \times H \times W}. \tag{5}$$

The DA module sequentially applies channel attention and spatial attention, followed by adaptive fusion, to refine F_n.

Channel attention emphasizes informative feature types while suppressing irrelevant or noisy channels. Global descriptors are obtained via average and max pooling across spatial dimensions:

$$F^c_{\text{avg}} = \text{AvgPool}(F_n), \quad F^c_{\text{max}} = \text{MaxPool}(F_n), \tag{6}$$

capturing both overall and extreme channel responses. These descriptors are processed through a two-layer Multi-Layer Perceptron (MLP) with reduction ratio r:

$$w_c = \sigma\Big(w_1 \cdot \text{ReLU}\big(w_0 \cdot (F^c_{\text{avg}} + F^c_{\text{max}})\big)\Big), \quad F_c = F_n \odot w_c, \tag{7}$$

where $w_0 \in \mathbb{R}^{C/r \times C}$ and $w_1 \in \mathbb{R}^{C \times C/r}$ are learnable weights, σ denotes sigmoid activation, and $\odot$ represents element-wise multiplication. This mechanism selectively emphasizes channels carrying critical edge or semantic cues.

Following channel refinement, spatial attention focuses on important regions within feature maps. Channel-refined features are aggregated along the channel dimension using average and max pooling,

$$F^s_{\text{avg}} = \text{AvgPool}(F_c), \quad F^s_{\text{max}} = \text{MaxPool}(F_c), \tag{8}$$

and are then concatenated and processed with a 5×5 convolution:

$$w_s = \sigma\Big(f_{5\times5}\big(\text{concat}(F^s_{\text{avg}}, F^s_{\text{max}})\big)\Big), \quad F_s = F_c \odot w_s. \tag{9}$$

This step captures spatial dependencies and highlights regions critical for detecting small, thin, or occluded objects.

Finally, channel-refined and spatial-refined features are adaptively fused:

$$F_p = \alpha \cdot F_c + \beta \cdot F_s, \quad \alpha + \beta = 1, \tag{10}$$

where α and β are learnable parameters balancing channel and spatial attention contributions.

The sequential design, which applies channel attention followed by spatial attention, leverages their complementary roles. Specifically, channel attention selects relevant feature types (e.g., edge or semantic cues), while spatial attention localizes critical regions. This enables the network to focus on fine structures and object boundaries. This proves particularly effective for 3D detection of small, sparse, or partially occluded objects.

By jointly modeling channel and spatial dependencies, the DA module enhances boundary delineation and improves occlusion robustness. Empirical evaluations demonstrate consistent improvements in precision and recall for small and occluded objects, validating its effectiveness in complex point cloud environments. Figure 3 illustrates the DA module's overall structure and workflow.

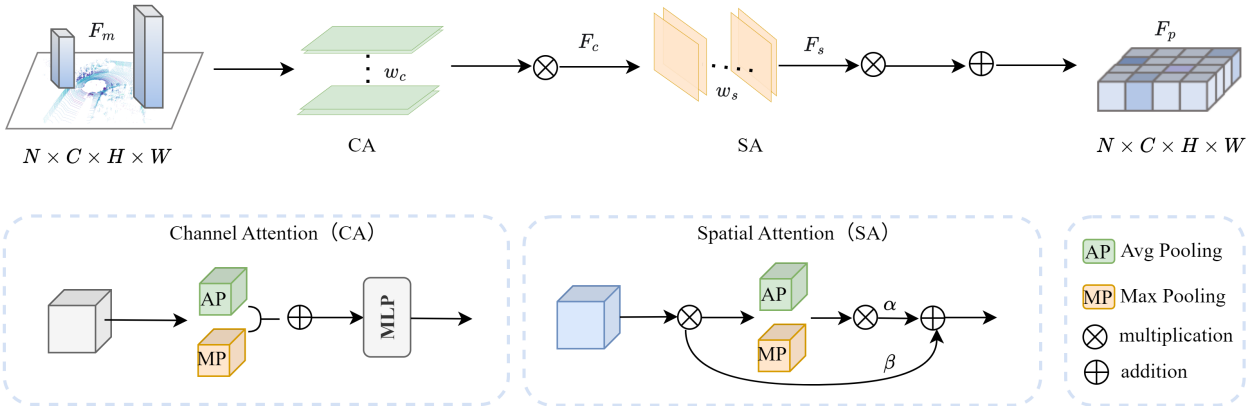

Figure 3. Architecture of the Dual Attention (DA) module, which refines feature representations via channel and spatial attention for precise boundary delineation and robust 3D object detection.

3.1.3. 3D Semantic Feature Extraction Module

The Semantic Feature (SF) module extracts rich contextual representations from 3D point clouds, preserving geometric fidelity while mitigating point sparsity. We adopt Cylinder3D [36], which transforms points from Cartesian coordinates (x, y, z) to cylindrical coordinates (ρ, θ, z). This transformation enables distance-adaptive voxelization, ensuring uniform sampling density across near and far regions and maintaining critical geometric structures for downstream detection.

An asymmetric 3D convolutional network with residual blocks is used to extract semantic features. Directional kernels (e.g., $1 \times 3 \times 3$, $3 \times 1 \times 3$) are employed to capture orientation-specific patterns, which are essential for objects with anisotropic shapes or

elongated structures, effectively handling the anisotropic nature of urban objects such as vehicles and pedestrians. A dimensional decomposition-based contextual module aggregates features along each axis, efficiently encoding global scene context without introducing significant computational overhead. The resulting semantic features complement the edge information extracted by the EF module, providing a unified representation that captures both local object boundaries and global semantic context. This design is especially effective for handling small, sparse, or partially occluded objects commonly found in remote sensing point clouds.

3.1.4. Feature Concatenation and Fusion

To integrate geometric, edge, and semantic cues, each point is assigned a semantic label and descriptor, aligned across datasets (e.g., KITTI, Waymo) via one-hot encoding. The final feature vector for each point is constructed as $[x_i, y_i, z_i, r_i, \phi_0, \phi_1, \phi_2, \phi_3]$, where (x_i, y_i, z_i) are the 3D coordinates, r_i represents the reflectance, and ϕ_j corresponds to the one-hot semantic encoding of the point.

Multi-scale alignment is applied to fuse fine-grained local edge details and global semantic context. This ensures that the network preserves critical geometric boundaries while integrating high-level semantic cues, enhancing robustness to sparse sampling, occlusion, and cluttered scenes. The fused feature representation thus provides a comprehensive point description, supporting reliable and accurate 3D object detection.

3.2. LiDAR-Based 3D Object Detection

The fused feature vectors from the SF, EF, and DA modules are input to a LiDAR-based detection head, which predicts oriented 3D bounding boxes represented by center coordinates, dimensions, orientation, and category. EAS-Det modifies only point-level features, leaving the detector architecture unchanged, ensuring full compatibility with both one-stage detectors and two-stage detectors. One-stage detectors directly regress box parameters in an anchor-free manner, simplifying the pipeline and improving efficiency. Two-stage detectors, on the other hand, generate coarse proposals and refine them using enriched features.

The network is trained with a multi-task loss that combines classification, 3D box regression, and orientation supervision. The classification loss guides category prediction, the regression loss ensures accurate box localization, and the orientation loss captures heading angles via sine-cosine or bin-based formulations. This ensures that the detector distinguishes object classes while precisely encoding geometric and directional properties, which is particularly beneficial for small, sparse, or partially occluded objects.

Extensive experiments on KITTI and Waymo demonstrate that EAS-Det consistently improves mean average precision across all categories, with notable gains for challenging classes such as pedestrians and cyclists. By leveraging the enriched feature representations, the framework achieves a favorable balance between detection accuracy, robustness to sparsity and occlusion, and computational efficiency. Its modular design allows seamless integration with existing LiDAR pipelines without additional overhead, making EAS-Det well-suited for real-world remote sensing scenarios.

4. Experiments and Results

This section presents the experimental setup and comprehensive evaluation of the proposed EAS-Det framework. We begin by introducing the datasets and implementation details, followed by performance comparisons with state-of-the-art 3D object detection methods. Finally, we conduct extensive ablation studies to validate the effectiveness and generalizability of the proposed modules across different detection architectures.

4.1. Implementation Details

4.1.1. Datasets and Evaluation Metrics

We evaluate the effectiveness of EAS-Det on two widely adopted benchmarks: the KITTI dataset and the Waymo Open Dataset. The KITTI dataset contains 7481 training samples and 7518 testing samples. Following the standard protocol, we use 3712 samples for training and 3769 for validation. Evaluation follows the KITTI 3D object detection benchmark, employing Average Precision (AP) and mean Average Precision (mAP) as primary metrics. The IoU thresholds are set to 0.7 for cars and 0.5 for pedestrians and cyclists. Detection difficulty is categorized into Easy, Moderate, and Hard levels based on occlusion and truncation degrees, which help assess model performance under different object conditions.

For semantic segmentation pre-training, we utilize the SemanticKITTI dataset, which provides point-wise annotations for 28 semantic classes, such as road, vegetation, and building. After consolidating dynamic and static object variants and filtering out underrepresented categories, we retain 19 classes. The semantic branch is pre-trained on this dataset, and we remap the 19-class annotations into three detection-relevant categories: car, pedestrian, and cyclist. This remapping employs a predefined rule-based one-hot encoding scheme, ensuring a deterministic correspondence between the segmentation labels and detection targets, which is crucial for seamless integration.

The Waymo Open Dataset features high-resolution LiDAR scans, with an average of 180,000 points per frame. The dataset is designed for autonomous driving applications and includes various real-world driving scenarios. We follow the official evaluation protocol, which includes two difficulty levels: LEVEL_1 (L_1) for nearby, minimally occluded objects, and LEVEL_2 (L_2) for distant or heavily occluded objects. Experiments are conducted under both settings to evaluate the model's cross-domain generalization capabilities in complex, dynamic driving environments.

4.1.2. Experimental Setup

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 2080 Ti GPU (11 GB VRAM), an Intel Core i7-10700K CPU, and 32 GB of system memory. The implementation is based on the PyTorch 1.11.0 deep learning framework with Python 3.8 and CUDA 11.3, and built upon the OpenPCDet framework [37]. The models are trained with a learning rate of 0.01 for the KITTI dataset and 0.003 for the Waymo dataset, using batch sizes of 8 and 16, respectively. We employ the OneCycleLR scheduler [38] to adjust the learning rate during training. Each model is trained for 80 epochs, with a weight decay of 0.01 and a momentum of 0.9.

4.1.3. Data Augmentation and Training Details

To enhance model robustness and reduce overfitting, we adopt comprehensive data augmentation strategies following established practices [5]. A sample database is constructed from the training set annotations, from which objects are randomly inserted during training. The augmentation pipeline includes random scaling, axis-aligned flipping, and rotation applied to both bounding boxes and point cloud data. These augmentations help simulate variations in object poses and orientations, enabling the model to generalize better to different object arrangements. For the Waymo dataset, additional techniques like range cropping and random downsampling are applied to accommodate its higher point density and larger sensing range. Range cropping focuses on sampling more points from the object regions of interest, while downsampling reduces computational complexity and ensures that the model can handle the high-density point clouds efficiently.

4.2. Comparison with State-of-the-Art

We evaluate EAS-Det on the KITTI and Waymo benchmarks, with hyperparameters optimized on respective validation sets. Comprehensive comparisons across different backbones, object categories, and difficulty levels demonstrate the effectiveness of our approach, with results summarized in Tables 1 and 2.

Table 1. Results on the KITTI test 3D detection benchmark for Cars, Pedestrians (Ped.), and Cyclists (Cyc.). The best and second-best performance values are highlighted in bold and underlined.

Method	Car AP_{3D}			Ped. AP_{3D}			Cyc. AP_{3D}		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
VoxelNeXt [22]	<u>89.85</u>	<u>81.92</u>	<u>78.45</u>	54.23	47.15	43.28	81.45	66.83	59.72
FusionPillars [39]	83.76	70.92	63.65	55.87	48.42	45.42	80.62	59.43	55.76
VoxelNextFusion [34]	**90.40**	**82.03**	**79.86**	52.56	45.72	41.85	79.28	64.74	58.25
Fast-CLOCs [35]	89.11	80.34	76.98	52.10	42.72	39.08	82.83	65.31	57.43
PointPainting [32]	82.11	71.70	67.08	50.32	40.97	37.87	77.63	63.78	55.89
SA-SSD [40]	88.75	79.79	74.16	-	-	-	-	-	-
SparseBEV [23]	89.23	81.05	77.86	55.41	48.33	44.26	82.01	67.39	<u>60.18</u>
PointPillars [8] †	79.05	74.99	68.30	52.08	43.53	41.49	75.78	59.07	52.92
BEVDet4D [41]	88.92	80.67	77.13	53.89	46.72	42.95	80.17	65.94	58.63
PointRCNN [9] †	85.94	75.76	68.32	49.43	41.78	38.63	73.93	59.60	53.59
SIENet [42]	88.22	81.71	77.22	-	-	-	<u>83.00</u>	<u>67.61</u>	60.09
SECOND [6]	83.13	73.66	66.20	51.07	42.56	37.29	70.51	53.85	46.90
EAS-Point (Ours)	88.52	78.45	77.60	**62.90**	**55.43**	**50.80**	**87.29**	**72.95**	**66.91**
EAS-Pillar (Ours)	86.61	77.44	75.99	<u>57.97</u>	<u>52.57</u>	<u>47.96</u>	78.14	61.73	58.12

† Baseline method.

Table 2. Performance comparisons on the Waymo Open Dataset validation set. Best and second-best values are in bold and underlined.

Method	Vehicle		Pedestrian		Cyclist	
	L_1	L_2	L_1	L_2	L_1	L_2
PointPillar [8] †	70.43	62.18	66.21	58.18	55.26	53.18
IA-SSD [29]	70.53	61.55	69.38	60.30	67.67	64.98
SECOND [6]	70.96	62.58	65.23	57.22	57.13	54.97
CenterPoint [28]	71.33	63.16	72.09	64.27	68.68	66.11
Part-A2-Anchor [43]	74.66	65.82	71.71	62.46	66.53	64.05
PV-RCNN [14] †	75.41	67.44	71.98	63.70	65.88	63.39
CenterPoint-Pillar [28]	70.50	62.18	<u>73.11</u>	<u>65.06</u>	65.44	62.98
Point-GNN [44]	75.26	<u>68.19</u>	65.86	61.14	68.69	<u>66.16</u>
FAR-Pillar [30]	71.30	63.02	67.15	58.90	58.26	56.06
GgPG [19]	**76.45**	**69.78**	66.72	62.32	**69.56**	**67.23**
EAS-PV (Ours)	<u>75.82</u>	68.07	**75.93**	**66.48**	<u>69.23</u>	65.48
EAS-Pillar (Ours)	73.60	65.25	71.61	63.22	63.07	61.28

† Baseline method.

4.2.1. On the KITTI Dataset

EAS-Det achieves consistent performance improvements across all categories and difficulty levels on the KITTI benchmark. As shown in Table 1, our framework delivers substantial AP gains over the best-performing multimodal method [34]: 10.34%, 9.71%, and 8.95% for pedestrians under easy, moderate, and hard settings, respectively, and 8.01%, 8.21%, and 8.66% for cyclists. These improvements underscore the effectiveness of the ESI module in capturing both fine-grained geometric details and global semantic context.

The results also reveal the architectural advantages of EAS-Det. EAS-Point excels in preserving structural details through point-based representations, which are critical for detecting small or occluded objects. In contrast, EAS-Pillar, while using efficient voxel-based

processing, achieves competitive performance with minimal computational overhead, demonstrating the complementary nature of our design across different backbone architectures.

4.2.2. On the Waymo Dataset

Evaluation on the large-scale Waymo Open Dataset validation set under challenging real-world conditions confirms the strong generalization capability of EAS-Det. EAS-Pillar improves LEVEL_1 Vehicle AP by 3.17 percentage points over the PointPillars baseline, while EAS-PV achieves competitive performance across all categories. Notably, EAS-PV establishes a new state-of-the-art in pedestrian detection on the validation set with 75.93% AP under LEVEL_1, significantly outperforming all compared methods.

These results underscore the robustness of our proposed modules, enabling consistent improvements across diverse sensing environments. The integration of edge geometry with semantic context via the ESI module ensures exceptional resilience in sparse and cluttered LiDAR scenes.

4.2.3. Visualization Analysis

To qualitatively demonstrate the operational effectiveness of our approach, we conduct visualization analyses on representative scenes from the KITTI dataset. Figure 4 presents a comparative visualization framework, contrasting detection results of PointRCNN and PointPillars with and without EAS-Det integration. The visualizations show that our method reliably detects diverse objects even under challenging conditions, including crowded scenes, background clutter, and partial occlusion, highlighting its robustness and real-world applicability.

Furthermore, the precision–recall (PR) curves in Figure 5 confirm that EAS-Det achieves a superior balance between precision and recall across both 3D and bird's-eye view (BEV) detection tasks. The curves maintain consistently higher precision across recall thresholds, validating the performance advantages of our edge-semantic fusion approach over the baseline detectors.

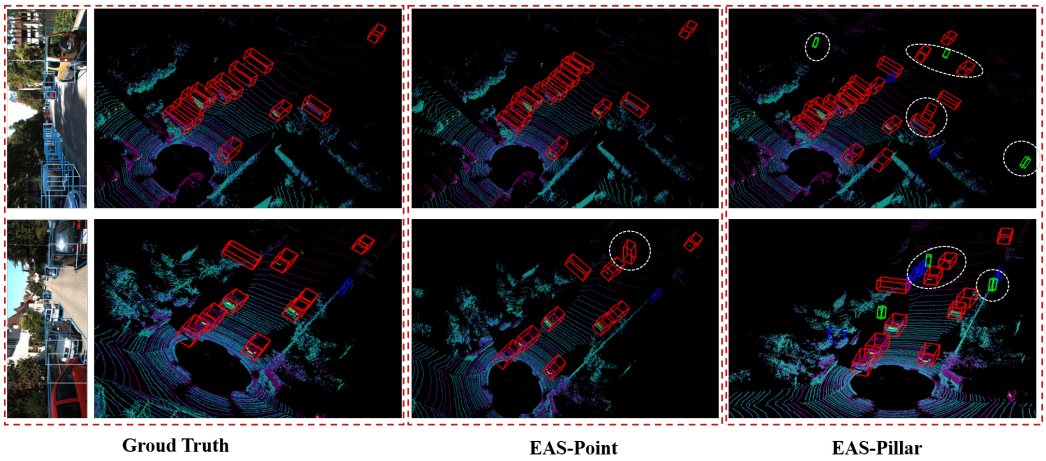

Figure 4. Visual analysis of detection results on the KITTI dataset. The left panel displays ground truth (GT) annotations, while the center and right panels present detection results from one-stage [8] and two-stage [9] methods, respectively. The scene contains multiple cars (red), cyclists (blue), and pedestrians (green), demonstrating performance under complex urban environments.

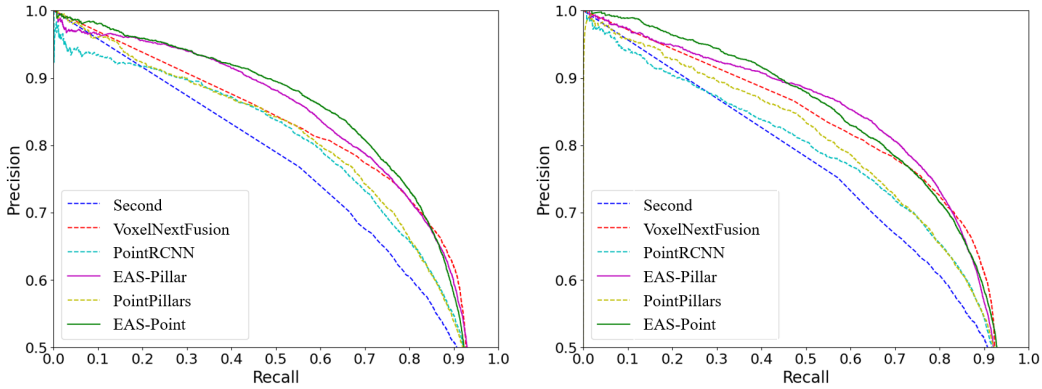

Figure 5. The Precision-Recall (PR) curves of mAP for 3D detection (**left**) and BEV detection (**right**).

4.3. Ablation Study

To comprehensively evaluate the contributions of each module and design choice, we conduct systematic ablation studies on the KITTI validation set. All variants are trained with identical hyperparameters, input resolutions, and optimization strategies to ensure fair comparison. Each configuration is trained independently until convergence, and performance metrics on the validation set are reported. In addition to accuracy, we analyze inference efficiency and model complexity to provide a holistic assessment.

4.3.1. Effectiveness of the Edge-Semantic Interaction Module

Table 3 presents the progressive integration of EF, DA, and ESI modules, highlighting their complementary contributions. The DA module yields substantial gains in 3D mAP, with improvements of +4.37% for EAS-Point and +1.28% for EAS-Pillar, while EF enhances geometric fidelity and boundary sharpness, providing 1.5–2.0% improvements across both backbones. These enhancements are critical for precise object contour modeling and bounding box alignment.

The synergistic effect of EF and DA is evident when combined: EAS-Point mAP increases from 63.42% to 66.39% with EF, indicating that EF reinforces geometric integrity while DA improves feature discriminability via channel and spatial attention. Together, they substantially enhance detection robustness and accuracy, particularly in complex scenes with small or partially occluded objects.

Table 3. Ablation and efficiency analysis of EAS-Det on the KITTI validation set. Each configuration incrementally integrates the DA, EF, and ESI modules to evaluate their individual and combined contributions. Metrics include 3D and BEV mAP, inference time, frame rate (FPS), FLOPs, and model size.

Method	Modules			Performance		Efficiency			
	ESI	EF	DA	mAP_{3D}	mAP_{BEV}	Time (ms)	FPS	FLOPs (G)	Size (MB)
EAS-Pillar				59.20	65.82	31	32	120.5	55.5
			✓	60.48	66.91	32	31	125.1	55.4
		✓	✓	62.60	69.34	37	27	128.7	55.6
	✓	✓	✓	63.91	70.23	42	24	130.2	55.7
EAS-Point				59.05	66.92	168	6	198.3	46.8
			✓	63.42	70.16	175	6	208.1	46.7
		✓	✓	66.39	72.83	188	5	215.9	46.9
	✓	✓	✓	68.94	74.15	212	5	221.5	47.0

4.3.2. Inference Speed and Model Size

We assess deployment efficiency across architectures as evidenced in Table 3. EAS-Pillar maintains real-time performance at 24–32 FPS with only 5–10 ms additional latency. In contrast, EAS-Point achieves higher accuracy with a 68.94% 3D mAP at 5–6 FPS, or 212 ms

per frame, while being suitable for offline high-precision tasks including HD mapping and autonomous driving log analysis.

All variants remain lightweight with parameter sizes below 56 MB, introducing minimal computational overhead. Compared to conventional LiDAR detectors such as SECOND and PointPillars, EAS-Det achieves a favorable balance between detection accuracy and efficiency, demonstrating scalability for both embedded and cloud-based deployments.

4.3.3. Fusion Strategy Analysis

We compare three feature fusion strategies, simple concatenation (Concat), element-wise addition (Add), and the proposed EAS mechanism, as summarized in Table 4. Concat and Add achieve 77.53% and 78.22% mAP, respectively, while EAS significantly outperforms both, achieving 80.91% mAP, including a +4.25% improvement in Pedestrian AP over Concat. This gain is attributed to EAS's adaptive weighting mechanism, which balances geometric and semantic information by suppressing redundant channels while emphasizing task-relevant features.

Table 4. Comparison of feature fusion strategies on the KITTI validation set. Three methods are evaluated: simple concatenation (Concat), element-wise addition (Add), and the proposed EAS mechanism. Metrics are 3D mAP values (%) for Car, Pedestrian, and Cyclist, along with overall mean AP.

Fusion Strategy	Car (AP_{3D})	Ped. (AP_{3D})	Cyc. (AP_{3D})/mAP
Concat	90.85	61.42	80.33/77.53
Add	91.02	62.18	81.47/78.22
EAS (Ours)	91.55	65.67	85.52/80.91

Unlike conventional fusion approaches that treat all features equally, the DA module within EAS performs adaptive feature re-weighting, which selectively enhances discriminative cues while attenuating less informative responses. By dynamically recalibrating feature activations according to both local geometric patterns and global semantic context, it enables a more synergistic integration of edge-aware and contextual information. This design proves particularly effective in LiDAR-based detection, where sparse and irregular point clouds require more structured feature interaction than simple concatenation or averaging can provide.

4.3.4. Robustness to Point Cloud Sparsity

To assess robustness under practical sensing conditions, we simulate point cloud sparsity by randomly downsampling input LiDAR data. Experiments use the EAS-Point configuration on the KITTI validation set under Moderate difficulty.

As shown in Table 5, the EF module exhibits notable stability: even at 25% point cloud density, the overall mAP decreases by a mere 2.96%, from 72.34% to 69.38%. This indicates that geometrically grounded edge cues remain informative despite severe sparsity.

Table 5. Robustness analysis of edge features under different point cloud densities on the KITTI validation set. The framework demonstrates consistent performance with moderate degradation under sparse conditions.

Sampling Ratio	Car (AP_{3D})	Ped. (AP_{3D})	Cyc. (AP_{3D})	mAP Δ
1.0 (full)	81.50	58.18	77.34	0.0
0.5 (Medium)	79.85	56.21	76.42	−1.65
0.25 (Low)	78.24	54.36	75.53	−3.12

Performance remains stable for categories with fewer points, such as pedestrians and cyclists, which are typically more challenging to detect. The edge features effectively capture salient contours and spatial boundaries, confirming that EAS-Det is accurate and reliable across varying point cloud densities, supporting its deployment in real-world autonomous perception systems.

4.3.5. Parameter Sensitivity Analysis

Following the commonly adopted practices in attention design [45], we conducted a comprehensive sensitivity analysis of the key hyperparameters in EAS-Det. Table 6 presents the effect of varying the Reduction Ratio (r) and Fusion Initialization Weights (α / β) on the 3D detection performance across the KITTI validation set. The Reduction Ratio notably influences performance. When varying r among {2, 4, 8}, the overall mAP remains within $\pm 0.5\%$, indicating that $r = 4$ is a robust and stable configuration. Similarly, the initialization of the fusion weights (α, β) has a negligible impact on the final performance. Only slight mAP variations (less than $\pm 0.3\%$) are observed when using alternative initializations (e.g., 0.2/0.8 or 0.8/0.2). These findings demonstrate that EAS-Det is largely insensitive to hyperparameter variations, ensuring ease of reproduction and adaptability across different settings.

Table 6. Parameter Sensitivity Analysis on the KITTI Validation Set. The effect of varying the Reduction Ratio (r) and Fusion Initialization Weights (α/β) on 3D detection performance.

Hyperparameter	Value	Car (AP_{3D})	Ped. (AP_{3D})	Cyc. (AP_{3D})
Reduction Ratio (r)	2	88.41	54.95	72.68
	4 (Default)	88.52	55.43	72.95
	8	88.35	55.10	72.70
Fusion Init. (α/β)	0.2/0.8	88.48	55.38	72.89
	0.5/0.5 (Default)	88.52	55.43	72.95
	0.8/0.2	88.45	55.35	72.86

5. Discussion

5.1. Module Contributions and Synergies

The ablation study demonstrates the distinct contributions and synergistic effects of each proposed component. The Edge Feature (EF) module enhances geometric representations, improving boundary localization critical for accurate object detection. The Dual Attention (DA) module adaptively balances spatial and channel dependencies, refining contextual relationships across multi-scale features. The Edge-Semantic Interaction (ESI) module optimally integrates structural and semantic cues, yielding the most significant performance gains, particularly for small or heavily occluded objects. These findings confirm the complementary nature of the proposed modules and their value as enhancements to mainstream 3D detection architectures.

A pivotal outcome of our experimental evaluation is the superior performance of EAS-Det in detecting small objects, namely pedestrians and cyclists. As quantitatively established in Section 4.2 (Table 1), our framework achieves substantial AP gains, up to +10.34% for pedestrians and +8.66% for cyclists on the KITTI benchmark, with consistent improvements across all difficulty levels. This demonstrates that our edge-semantic fusion approach is particularly effective in capturing the fine-grained structural details that are critical for recognizing objects with minimal point cloud signatures.

5.2. Architectural Compatibility and Performance

EAS-Det demonstrates consistent performance improvements across both single-stage and two-stage detection paradigms. Single-stage detectors maintain real-time performance (24–32 FPS) while benefiting from edge-aware feature fusion, showing particular effectiveness for medium- and large-scale objects. Two-stage detectors achieve higher precision in fine-grained localization, with the ESI modules significantly enhancing accuracy for small or partially occluded objects. Notably, EAS-Det improves detection across all object categories and difficulty levels, with the most substantial gains observed for challenging cases involving small or occluded objects. This improvement stems from the framework's unique ability to preserve object boundaries and enhance feature alignment through explicit geometric-semantic interactions—a capability absent in prior fusion-based approaches.

5.3. Robustness Under Sparse and Noisy Conditions

EAS-Det demonstrates strong robustness to sparse and noisy LiDAR data, a key requirement for real-world deployment. Its performance remains stable even at 25% of the original point density, as edge-aware features effectively preserve structural information under sparsity. The EF module's gradient-based edge extraction captures key geometric boundaries, ensuring reliable detection in sparse or long-range sensing. Compared to density- or semantics-only methods, our edge-semantic fusion provides superior resilience in challenging perceptual conditions.

The framework also handles noisy inputs effectively. Gradient-based edge features emphasize structural variations while suppressing local noise during BEV voxelization and convolution, preserving clear object contours. In sparsity experiments (Table 5), EAS-Det remains competitive even with 50% of points missing or corrupted. Improved performance on challenging KITTI and Waymo scenes, under occlusion, clutter, and sensor noise, further validates its robustness and practical viability for real-world LiDAR perception systems.

5.4. Computational Efficiency and Deployment

EAS-Det achieves an optimal balance between detection accuracy and computational efficiency. As summarized in Table 3, the complete EAS-Pillar model improves the 3D mAP by +4.71% over the baseline with only an 8.1% increase in FLOPs (from 120.5 G to 130.2 G) and an almost unchanged model size of 55.7 MB, achieving real-time inference at 24 FPS. Similarly, EAS-Point attains a +9.89% mAP gain while introducing 11.7% additional FLOPs and maintaining a compact 47.0 MB model.

The modular architecture allows seamless integration into mainstream backbones without full retraining, enabling flexible adaptation to diverse sensing configurations. These results demonstrate that EAS-Det delivers substantial performance improvements with minimal computational and memory overhead, ensuring suitability for both embedded and edge-computing platforms. This strong efficiency–accuracy trade-off further underscores its potential for safety-critical, real-time applications such as autonomous driving and remote sensing.

5.5. Qualitative Analysis

To visually validate the performance of EAS-Det, particularly its superior capability in small-object detection, we provide a qualitative comparison as shown in Figure 6. Empowered by edge-aware semantic fusion, our framework exhibits remarkable robustness in detecting challenging instances such as distant pedestrians and cyclists, where baseline detectors (e.g., PointPillars) often fail to generate bounding boxes due to sparse point distributions and weak semantic features. This qualitative evidence directly correlates with the significant quantitative improvements for pedestrians and cyclists reported in

Table 1, collectively substantiating that our approach effectively enhances the perception of small-scale targets in complex point clouds.

(a) Long-range scene (b) Small-object scene (c) Crowded urban scene

Figure 6. Empowered by the Edge-Semantic Interaction (ESI) module, the proposed method achieves more reliable detections and sharper boundaries for small-object detection.

5.6. Implications for Multi-Platform Sensing

While validated on ground-based LiDAR datasets, EAS-Det's core principles address a key challenge in remote sensing—the effective fusion of structural and contextual cues across heterogeneous platforms such as satellite, aerial, and ground-based systems. The proposed edge-semantic interaction mechanism provides a unified solution, forming the foundation for cross-domain perception models that integrate multi-source sensing perspectives. Insights gained from autonomous driving applications can further drive advancements in large-scale 3D scene understanding and collaborative sensing, contributing to more efficient and generalizable remote sensing systems.

6. Limitations and Future Work

Despite its advantages, EAS-Det exhibits several limitations that suggest directions for future research.

First, the current implementation relies solely on LiDAR data. While LiDAR is robust in low-light and adverse weather, its performance degrades under extreme sparsity, during long-range perception or heavy precipitation. Future work will explore multi-modal fusion with RGB cameras, radar, and event-based sensors to develop adaptive strategies for selecting the most informative cues under varying conditions.

Second, although EAS-Det is architecturally extendable to multi-modal fusion, practical implementation faces challenges including spatiotemporal calibration, feature alignment, and modality conflicts. Future work will develop robust fusion frameworks incorporating uncertainty-aware weighting and self-supervised alignment strategies to prevent performance degradation.

Third, the model's decision-making process remains largely opaque despite some interpretability provided by edge and semantic visualizations. This black-box nature poses challenges for debugging and safety certification. Future research will integrate explainable AI techniques, such as attention map visualization, saliency-based feature attribution, and uncertainty quantification, to enhance transparency and facilitate failure diagnosis.

Fourth, EAS-Det currently lacks explicit modeling of predictive uncertainty, which is crucial for safety-critical tasks. Future work will integrate Bayesian deep learning and ensemble methods to estimate calibrated confidence measures, enabling reliable decision-making under domain shifts and sensor degradation.

Finally, while EAS-Det achieves acceptable efficiency on high-end GPUs, its deployment on resource-constrained embedded systems remains unverified. Future efforts will focus on model optimization through compression, pruning, quantization, and knowledge distillation. For specific application scenarios such as drones and autonomous vehicles, we will extend evaluation to large-scale, cross-domain datasets to ensure robustness in challenging real-world environments.

7. Conclusions

This paper presents EAS-Det, a novel LiDAR-based 3D object detection framework that explicitly integrates semantic context, geometric structure, and edge information through a dual-attention mechanism. By enabling multi-scale feature fusion, EAS-Det produces highly discriminative and robust representations. Extensive experiments on the KITTI and Waymo datasets demonstrate that EAS-Det consistently outperforms baseline methods across both single-stage and two-stage architectures, exhibiting remarkable robustness in detecting small and challenging objects such as pedestrians and cyclists under dense traffic and sparse point conditions.

The proposed approach enhances detection accuracy without sacrificing efficiency, while maintaining high inference speed and a compact footprint suitable for deployment. Future work will focus on extending EAS-Det to multi-modal detection by incorporating RGB and radar inputs, and on further optimizing its implementation for embedded and edge-computing platforms. These efforts will pave the way for unified cross-domain perception in autonomous systems.

Author Contributions: Conceptualization, H.W. and Y.Z.; methodology, software, formal analysis, and visualization, H.W.; validation, J.Z. and F.C.; writing—original draft preparation, H.W. and Y.Z.; writing—review and editing, H.W. and J.Z.; supervision, project administration, and funding acquisition, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets analyzed during the current study are publicly available. The KITTI dataset can be accessed at http://www.cvlibs.net/datasets/kitti/ (accessed on 20 March 2012), and the Waymo Open Dataset can be accessed at https://waymo.com/open/ (accessed on 20 April 2021).

Acknowledgments: The authors thank the authors of "SeSame: Simple, Easy 3D Object Detection with Point-Wise Semantics" [33] for sharing their dataset partition details. We are also grateful to the OpenPCDet Laboratory for the computational resources and to the anonymous reviewers for their insightful and constructive feedback. The authors take full responsibility for the content of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

FPS	Farthest Point Sampling
UAV	Unmanned Aerial Vehicle
BEV	Bird's-Eye View
AP	Average Precision
mAP	Mean Average Precision
IoU	Intersection over Union
CNN	Convolutional Neural Network
FPN	Feature Pyramid Network
RCNN	Region-based Convolutional Neural Network
RPN	Region Proposal Network
NMS	Non-Maximum Suppression

References

1. Jiang, K.; Huang, J.; Xie, W.; Lei, J.; Li, Y.; Shao, L.; Lu, S. DA-BEV: Unsupervised Domain Adaptation for Bird's Eye View Perception. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 322–341.
2. Feng, T.; Wang, W.; Ma, F.; Yang, Y. LSK3DNet: Towards Effective and Efficient 3D Perception with Large Sparse Kernels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–18 June 2024; pp. 14916–14927.
3. Xie, Y.; Xu, C.; Rakotosaona, M.-J.; Rim, P.; Tombari, F.; Keutzer, K.; Tomizuka, M.; Zhan, W. SparseFusion: Fusing Multi-Modal Sparse Representations for Multi-Sensor 3D Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 17591–17602.
4. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 77–85.
5. Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4490–4499.
6. Yan, Y.; Mao, Y.; Li, B. SECOND: Sparsely Embedded Convolutional Detection. *Sensors* **2018**, *18*, 3337. [CrossRef] [PubMed]
7. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-View 3D Object Detection Network for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6526–6534.
8. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. PointPillars: Fast Encoders for Object Detection from Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 12689–12697.
9. Shi, S.; Wang, X.; Li, H. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 770–779.
10. Sun, P.; Wang, W.; Chai, Y.; Elsayed, G.; Bewley, A.; Zhang, X.; Sminchisescu, C.; Anguelov, D. RSN: Range Sparse Net for Efficient, Accurate LiDAR 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 5721–5730.
11. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6568–6577.
12. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
13. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2446–2454.
14. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10529–10538.
15. Virasova, A.Y.; Klimov, D.I.; Khromov, O.E.; Gubaidullin, I.R.; Oreshko, V.V. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Radioengineering* **2021**, *30*, 115–126. [CrossRef]
16. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
18. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef] [PubMed]
19. Wang, K.; Zhou, M.; Lin, Q.; Niu, G.; Zhang, X. Geometry-Guided Point Generation for 3D Object Detection. *IEEE Signal Process. Lett.* **2025**, *32*, 136–140. [CrossRef]
20. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. AVOD: Aggregate View Object Detection in Autonomous Driving. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1852–1859.
21. Qiao, S.; Loy, C.C.; Yuille, A. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June2021; pp. 10213–10224.

22. Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; Jia, J. VoxelNeXt: Fully Sparse VoxelNet for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 21674–21683.
23. Liu, H.; Teng, Y.; Lu, T.; Wang, H.; Wang, L. SparseBEV: High-Performance Sparse 3D Object Detection from Multi-Camera Videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 18580–18590.
24. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 1256–1270.
25. Li, H.; Zhang, Q.; Liu, Y.; Wang, J. YOLO-RS: A Lightweight YOLO Architecture for High-Resolution Remote Sensing Image Object Detection. *Remote Sens.* **2024**, *16*, 1456.
26. Zhao, K.; Zhang, R.; Zhou, Y.; Chen, L. Edge-YOLO: An Edge-Device Oriented Real-Time Object Detector with Hardware-Aware Neural Architecture Search. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 5689–5703.
27. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. *Lecture Notes Comput. Sci.* **2016**, *9905*, 21–37.
28. Yin, T.; Zhou, X.; Krähenbühl, P. Center-Based 3D Object Detection and Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11784–11793.
29. Zhang, C.; Liu, Z.; Chen, X.; Li, J.; Wang, H.; Wu, B.; Zhou, S.; Yang, L.; Zhao, Q.; Xu, K.; et al. IA-SSD: Multi-Scale Information Fusion and Adaptive Attention for 3D Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual Conference, 11–17 October 2021; pp. 15667–15676.
30. Wang, W.; Ma, C.; Zhang, Y.; Li, X.; Chen, Z.; Liu, Y.; Yang, H.; Zhao, R.; Wu, S.; Zhou, T.; et al. FAR-Net: Feature Adaptive Refinement Network for Oriented 3D Object Detection. *IEEE Trans. Intell. Transp. Syst.* **2023**. [CrossRef]
31. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum PointNets for 3D Object Detection from RGB-D Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 918–927.
32. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. PointPainting: Sequential Fusion for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4603–4611.
33. Oh, H.; Yang, C.; Huh, K. SeSame: Simple, Easy 3D Object Detection with Point-Wise Semantics. In Proceedings of the Asian Conference on Computer Vision (ACCV), Hanoi, Vietnam, 8–12 December 2024; pp. 211–227.
34. Song, Z.; Zhang, G.; Xie, J.; Liu, L.; Jia, C. VoxelNextFusion: A Simple, Unified and Effective Voxel Fusion Framework for Multi-Modal 3D Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12.
35. Pang, S.; Morris, D.; Radha, H. Fast-CLOCs: Fast Camera-LiDAR Object Candidates Fusion for 3D Object Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 187–196.
36. Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Ma, Y.; Li, W.; Li, H.; Lin, D. Cylindrical and Asymmetrical 3D Convolution Networks for LiDAR Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June2020; pp. 9934–9943.
37. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H.; Yang, R.; Zhang, S.; Wu, Y.; et al. OpenPCDet: An Open-Source Toolbox for 3D Object Detection from Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
38. Smith, L.N. A Disciplined Approach to Neural Network Hyper-Parameters: Part 1—Learning Rate, Batch Size, Momentum, and Weight Decay. *arXiv* **2018**, arXiv:1803.09820.
39. Zhang, J.; Xu, D.; Li, Y.; Zhao, L.; Su, R. FusionPillars: A 3D Object Detection Network with Cross-Fusion and Self-Fusion. *Remote Sens.* **2023**, *15*, 2692. [CrossRef]
40. He, C.; Zeng, H.; Huang, J.; Hua, X.-S.; Zhang, L. Structure Aware Single-Stage 3D Object Detection from Point Cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11870–11879.
41. Huang, J.; Huang, G. BEVDet4D: Exploit Temporal Cues in Multi-camera 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 1901–1910.
42. Li, Z.; Yao, Y.; Quan, Z.; Xie, J.; Yang, W. Spatial Information Enhancement Network for 3D Object Detection from Point Cloud. *Pattern Recognit.* **2022**, *128*, 108692. [CrossRef]
43. Shi, S.; Wang, Z.; Li, X.; Hou, L.; Guo, C.; Ma, C.; Li, H. Part-A2 Net: 3D Part-Aware and Aggregation Neural Network for Object Detection from Point Cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 820–829. [CrossRef]

44. Shi, W.; Rajkumar, R. Point-GNN: Graph Neural Network for 3D Object Detection in a Point Cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 1708–1716.
45. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.