information

Article

Boosting LiDAR Point Cloud Object Detection via Global Feature Fusion

Xu Zhang, Fengchang Tian, Jiaxing Sun and Yan Liu *

School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450000, China; zhangxu@zzuli.edu.cn (X.Z.); 332405080490@zzuli.edu.cn (F.T.); 332207010619@zzuli.edu.cn (J.S.)
* Correspondence: lyanzju@zzuli.edu.cn

Abstract

To address the limitation of receptive fields caused by the use of local convolutions in current point cloud object detection methods, this paper proposes a LiDAR point cloud object detection algorithm that integrates global features. The proposed method employs a Voxel Mapping Block (VMB) and a Global Feature Extraction Block (GFEB) to convert the point cloud data into a one-dimensional long sequence. It then utilizes non-local convolutions to model the entire voxelized point cloud and incorporate global contextual information, thereby enhancing the network's receptive field and its capability to extract and learn global features. Furthermore, a Voxel Channel Feature Extraction (VCFE) module is designed to capture local spatial information by associating features across different channels, effectively mitigating the spatial information loss introduced during the one-dimensional transformation. The experimental results demonstrate that, compared with state-of-the-art methods, the proposed approach improves the average precision of vehicle, pedestrian, and cyclist targets on the Waymo subset by 0.64%, 0.71%, and 0.66%, respectively. On the nuScenes dataset, the detection accuracy for var targets increased by 0.7%, with NDS and mAP improving by 0.3% and 0.5%, respectively. In particular, the method exhibits outstanding performance in small object detection, significantly enhancing the overall accuracy of point cloud object detection.

Keywords: laser point cloud; 3D; global feature; SSM

check for updates

Academic Editor: Heming Jia

Received: 25 July 2025
Revised: 19 September 2025
Accepted: 24 September 2025
Published: 26 September 2025

Citation: Zhang, X.; Tian, F.; Sun, J.; Liu, Y. Boosting LiDAR Point Cloud Object Detection via Global Feature Fusion. *Information* **2025**, *16*, 832. https://doi.org/10.3390/info16100832

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 burgeoning field of autonomous driving technology has positioned LiDAR (Light Detection and Ranging) point cloud object detection as a pivotal research domain within its core perception stack. LiDAR sensors, by virtue of their capacity to directly acquire precise spatial coordinates and geometric configurations of objects in three-dimensional space, furnish autonomous driving systems with indispensable, high-fidelity environmental perception inputs. However, in contrast to the structured and uniformly dense characteristics of image data, point cloud data presents inherent challenges that stem from its sparsity, irregular distribution, and unordered nature. These intrinsic properties render the direct application of conventional 2D image processing and detection paradigms largely unsuitable for point cloud object detection tasks. Consequently, the development of highly efficient methodologies for extracting discriminative feature representations from high-dimensional, non-Euclidean point cloud data, thereby enabling accurate and robust object detection, represents a critical scientific challenge currently confronting the autonomous driving perception community.

Existing methodologies for point cloud object detection can be broadly categorized into three principal types, differentiated by their data representation and feature extraction strategies: point-based methods, voxel-based methods, and point–voxel fusion approaches. Point-based methods operate directly on the raw, unordered point sets, thereby maximally preserving the fine-grained geometric topology and intricate spatial details of the original data. Nevertheless, these methods typically entail substantial computational complexity and memory overhead and may encounter performance bottlenecks when confronted with extremely sparse point clouds. Voxel-based methods, conversely, transform irregular point clouds into uniform three-dimensional voxel grids, thereby facilitating the convenient application of mature 3D convolutional neural networks for feature learning [1,2]. However, this transformation process invariably introduces quantization errors and can lead to the loss of granular spatial information, particularly when coarser voxel resolutions are employed. To judiciously leverage the strengths of both paradigms while mitigating their respective limitations, point–voxel fusion methods have emerged, aiming to achieve superior detection performance through the synergistic learning of multi-modal features. While notable progress has been attained across these various methodologies, a prevalent characteristic is their reliance on local receptive field convolutional operations for feature extraction. This inherent constraint on the receptive field impedes the network's ability to effectively capture global contextual information within point cloud data. This limitation is particularly pronounced in the detection of physically smaller targets, such as pedestrians and cyclists, where an insufficient receptive field severely compromises detection efficacy [3]. Consequently, the development of mechanisms capable of effectively expanding the network's receptive field and extracting global features from point cloud data is of paramount importance for elevating the overall performance of point cloud object detection networks.

To surmount the aforementioned challenges inherent in localized feature extraction, this study draws inspiration from the groundbreaking advancements in sequence modeling, particularly the state space model (SSM) and its sophisticated derivative, the Mamba architecture [4]. The core strength of the Mamba model lies in its unique selective state space model mechanism, which, through the introduction of an innovative discretization process, enables the seamless integration of SSM into deep learning frameworks. This mechanism endows the Mamba model with exceptional capabilities for processing long sequence data, allowing it to efficiently capture global dependencies with approximately linear computational complexity. Inspired by this cutting-edge paradigm, this paper proposes a novel LiDAR point cloud object detection method that ingeniously integrates global features. This approach transforms the raw point cloud data into a one-dimensional long sequence and performs highly efficient global feature extraction on this sequence. This strategy fundamentally aims to overcome the receptive field limitations of conventional local convolutions, significantly augmenting the network's global perception and modeling capacities, thereby substantially improving the detection accuracy and robustness, especially for small-sized targets within complex autonomous driving scenarios.

2. Materials and Methods

2.1. Point Cloud-Based Target Detection Methods

Point-based detection methods, which learn geometric features directly from irregular and disordered point clouds, have advanced through several key contributions. For instance, Qi et al. [5] introduced PointNet++, a hierarchical neural network that captures fine-grained geometric features through recursive multi-scale feature learning, enabling robust processing of unstructured 3D data. Choy et al. [6] proposed Minkowski convolutional neural networks, extending convolutional operations to sparse, high-dimensional spatio-

temporal point clouds for efficient 4D data processing in detection tasks. Deng et al. [7] proposed PoIFusion, a multi-modal 3D object detection framework that fuses RGB images and LiDAR point clouds at points of interest (PoIs), achieving efficient and accurate detection without relying on global attention mechanisms. Zhang et al. [8] introduced 3DGeoDet, a geometry-aware image-based 3D object detection method that leverages explicit and implicit 3D geometric representations to enhance detection robustness and generality. Hamilton [9] developed foundational state space models for time series analysis, adaptable to sequential point cloud data, while Gu and Dao [4] proposed Mamba, a linear time sequence modeling approach with selective state spaces for efficient point cloud processing. These point-based approaches often leverage PointNet and its variants, such as 3DSSD by Yang et al. [10], which achieves efficient single-stage 3D object detection, or the method by Zhang et al. [11], which prioritizes key points for highly efficient LiDAR-based detectors. For example, Shi et al. [12] proposed segmenting foreground points to generate and refine target proposals in a standard coordinate system for accurate detection.

In contrast, voxel-based methods convert irregular point clouds into regular voxel representations to enhance network learning efficiency. Graham et al. [13,14] developed submanifold sparse convolutional networks, optimizing sparse 3D data processing for semantic segmentation and detection. Lang et al. [15] introduced PointPillars, a fast encoder transforming point clouds into 2D pseudo-images for efficient detection. Yan et al. [16] proposed SECOND, a sparsely embedded convolutional method, while Zheng et al. [17] presented CIA-SSD, a confident IoU-aware single-stage detector for enhanced precision. Wang et al. [18] introduced Voxel-RCNN-Complex for complex traffic conditions, Liu et al. [19] proposed a point attention mechanism for voxel-based detection, Chen et al. [20] developed VoxelNeXt for fully sparse end-to-end detection and tracking, An et al. [21] introduced SP-Det using saliency prediction for sparse point clouds, and Yin et al. [22] proposed a center-based approach to detection and tracking. Recent advancements in point–voxel fusion methods combine point and voxel features to improve accuracy. Wang et al. [23] proposed DSVT, a dynamic sparse voxel Transformer using rotated sets. Shi et al. [24] introduced PV-RCNN, employing Farthest Point Sampling for feature extraction, with PV-RCNN++ [25] optimizing sampling and grid pooling for faster processing. Sheng et al. [26] used a Transformer-based region of interest (ROI) approach, though it was limited by receptive field constraints. Liu et al. [27] developed a point–voxel fusion network, Yang et al. [28] proposed PVT-SSD with a point–voxel Transformer, and Deng et al. [29] introduced PVC-SSD with dual-channel fusion. Tian et al. [30] applied fusion for irrigation system identification, Liu et al. [31] developed PVA-GCN for 3D pose estimation, Xu et al. [32] proposed self-supervised point cloud pre-training, and the authors of [33] provided SpConv, a sparse convolution library for efficient voxel processing. Thus, extending the network's receptive field and enhancing global modeling remain critical for effective point cloud target detection.

2.2. State Space Models

In recent years, state space models (SSMs) have gained significant attention for their ability to model sequential data, leading to several impactful studies. Gu et al. [34] introduced structured state spaces to efficiently model long sequences, providing a foundation for scalable sequence processing. Gu et al. [35] combined recurrent, convolutional, and continuous-time models with linear state space layers, enhancing sequence modeling flexibility. Smith et al. [36] proposed simplified state space layers to streamline sequence modeling with reduced computational complexity. Liang et al. [37] extended SSMs to the vision domain with Vision Mamba, demonstrating their generality in visual representation learning. Liu et al. [38] developed VMamba, incorporating a cross-scanning module to

selectively scan 2D images, and they achieved notable results in image classification. The unordered and sparse nature of point cloud data poses a significant challenge for global feature extraction in 3D object detection. Traditional convolutional methods, constrained by local receptive fields, struggle to effectively capture long-range spatial dependencies in point clouds. In contrast, the selective state space model (SSM), particularly the Mamba architecture, leverages its dynamic state update mechanism to process long sequences with near-linear computational complexity (O(n)), making it particularly well-suited for voxelized point cloud sequence representations. Mamba's discretization process transforms continuous states into deep learning-compatible representations, enabling adaptive focus on critical spatial features and enhancing global context modeling. Compared to Transformer-based models, Mamba avoids the quadratic complexity ($O(n^2)$) of attention mechanisms while retaining the ability to model long-range dependencies. This provides a theoretical advantage in handling the high-dimensional, non-Euclidean structure of point clouds, significantly improving network robustness and accuracy, especially for small object detection (e.g., pedestrians and cyclists) in sparse point cloud scenarios. Yu et al. [39] showed through extensive experiments that, while SSMs are not essential for image classification, their ability to extract global features from long data sequences is highly valuable for detection and segmentation tasks. Inspired by this, enhancing the network's receptive field through global feature extraction from point cloud data is critical for improving point cloud target detection performance. To this end, this paper proposes a novel LiDAR point cloud object detection with a global feature (LDGF) method, which processes point cloud data through a Point Cloud Feature Extract Block (PFEB) to capture global features. Within the PFEB, a Voxel Mapping Block (VMB) maps point cloud data into a one-dimensional long sequence, followed by a Global Feature Extract Block (GFEB) that extracts global features. Additionally, a Voxel Channel Feature Extract (VCFE) module is introduced to enhance local feature extraction, mitigating the loss of spatial information due to data one-dimensionalization. Experiments on the Waymo-mini and nuScenes datasets demonstrate that LDGF achieves effective point cloud target detection and robust performance.

2.3. Network Architecture

The overall network architecture of LDGF is shown in Figure 1. The point cloud data is first subjected to a Voxelization operation [40], which splits the 3D space into voxel blocks of equal size Fvoxel.

LDGF then uses PFEB to extract global features of the point cloud data from the voxel blocks. Specifically, the Fvoxel in PFEB will be mapped into a one-dimensional long data sequence by the VMB, and the global features will be extracted by GFEB. The VCFE module supplements the missing spatial information due to the one-dimensionalization of the point cloud data by associating different channel features to enhance the network's ability to learn and detect point cloud targets. Finally, the network obtains the detection results with the BEV Backbone and Detection Head.

2.4. Point Cloud Feature Extract Block

As shown in Figure 2, the VMB traverses the voxel blocks based on the coordinates, and non-empty voxels are recorded during the traversal process and output as a mapped point cloud data sequence after the traversal is completed.

Assuming that the voxel coordinates are (xm, ym, zm), where m $\in$ [0,1,...,n], n is the number of non-empty voxels, and the traversal starts from the voxel block with the coordinates of (0, 0, 0), during the process, the VMB first traverses the voxel block that is at the X-axis coordinate of 0 and traverses along the Y-axis positively in this two-dimensional plane from the bottom (with small Z-axis coordinates) to the top (with large

Z-axis coordinates) traversal (direction of the blue arrow in the scan orange voxel block in Figure 2). Subsequently, the VMB continues to traverse the remaining voxel blocks in the X-axis direction as described above (blue arrow direction in the upper left corner of the scan in Figure 2). This traversal recording method not only maps the point cloud data into a one-dimensional sequence but also maintains the spatial correlation between neighboring voxels to some extent.

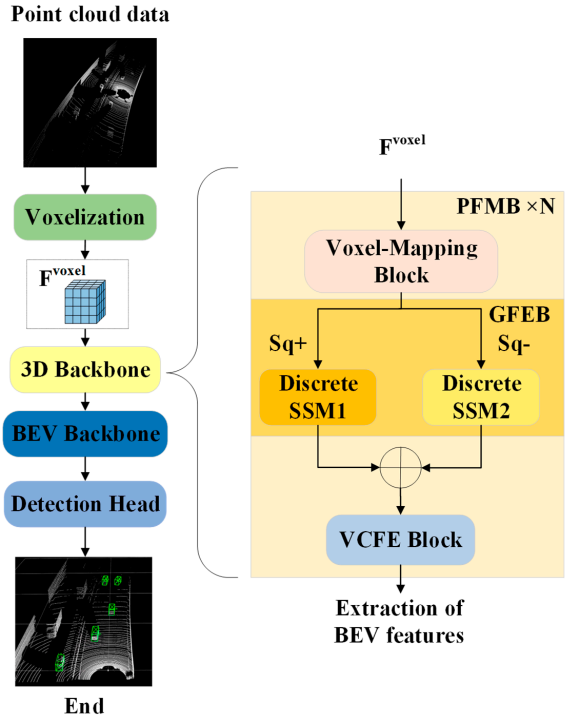


Figure 1. LDGF network architecture diagram.

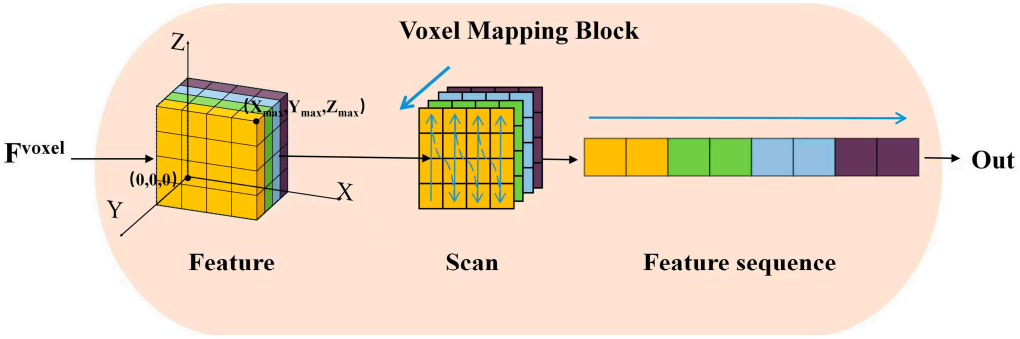


Figure 2. The mapping method of voxels in the VMB.

To reduce the effect of traversal order on point cloud feature extraction, the sequence Sq+ generated after VMB mapping is inverted and processed to obtain the sequence Sq−, which is in the opposite direction of the original sequence.

The VMB adopts a coordinate-based traversal order (starting from X = 0, traversing positively along the Y and Z axes) to ensure spatial continuity in the voxel serialization process (see Algorithm 1 and Figure 2). This order mimics the spatial structure of point clouds, facilitating the preservation of local correlations among neighboring voxels. However, a unidirectional traversal order may introduce directional bias, where feature extraction becomes dependent on the specific traversal path, leading to uneven modeling of the point cloud's spatial structure. For instance, in a unidirectional traversal, voxels near the sequence's starting point (e.g., near X = 0) may receive higher weights during feature aggre-

gation, while those further along (e.g., at the end of the X-axis) may be underrepresented. This bias could compromise the completeness of global feature extraction, particularly in sparse point clouds, where certain regions may be suboptimally processed due to the serialization order. To mitigate this potential impact, we introduce a bidirectional processing mechanism using forward (Sq+) and reverse (Sq−) sequences (see Algorithm 1 and Figure 2). Theoretically, bidirectional processing balances the directional bias by fusing feature representations from opposing traversal directions. Specifically, the forward sequence prioritizes spatial relationships from the X-axis start to end, while the reverse sequence does the opposite. The fusion of these features, facilitated by the SSM's dynamic state update mechanism, enhances the model's robustness to the point cloud's global structure. This bidirectional strategy is theoretically equivalent to symmetrizing the spatial representation of the serialized point cloud, thereby reducing feature loss or bias caused by unidirectional traversal and ensuring the completeness and consistency of global feature extraction.

GFEB extracts the voxel features from the Sq+ and Sq− sequences using two discretized SSMs with the same structure. After fusing the two extracted features, they are reduced to the same size as the input voxel feature size for output. This process enriches the information of feature extraction, compensates for the deficiency of the traditional local sensing field methods, ensures the point cloud global feature extraction completeness, and improves the adaptability of the network to detect small targets.

Algorithm 1: Bidirectional Voxel Mapping Block (VMB)

```
Input:
Voxel feature tensor F_voxel ∈ R^(X_max × Y_max × Z_max × C)
Output:
Serialized bidirectional feature sequence F_out
Initialize empty sequences Sq+ ← [], Sq− ← []
# Forward traversal (Sq+)
for x = 0 to X_max−1 do
    for y = 0 to Y_max−1 do
        for z = 0 to Z_max−1 do
            Sq+.append(F_voxel [x, y, z, :])
        end for
    end for
end for
# Reverse traversal (Sq−)
for x = X_max−1 downto 0 do
    for y = Y_max−1 downto 0 do
        for z = Z_max−1 downto 0 do
            Sq−.append(F_voxel [x, y, z, :])
        end for
    end for
end for
# Feature extraction using two SSMs
F+ ← SSM(Sq+)
F− ← SSM(Sq−)
# Fusion and dimensionality reduction
F_fused ← Fuse(F+, F−)
F_out ← Reduce(F_fused, target_dim = C)
return F_out
```

2.5. Voxel Channel Feature Extract

The overall network architecture of LDGF is shown in Figure 1. The point cloud data is first subjected to a voxelization operation, which splits the 3D space into voxel blocks of equal size F^{voxel}.

One-dimensionalization of point cloud data leads to the loss of local spatial feature information. For this reason, this paper proposes the VCFE module to strengthen the network's ability to capture and extract local spatial features. Figure 3 shows the structure of the VCFE module. For a given input feature *FL*, VCFE first divides the channels and convolves each channel individually to ensure that the information from other channels will not be mixed in. As a result, the network can effectively extract the local spatial information of the input feature map.

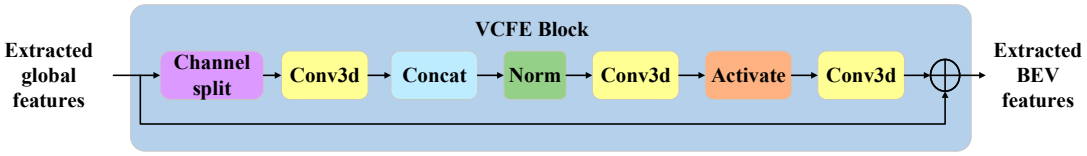


Figure 3. Detailed views of the VCFE module.

For a given input feature *FL*, the VCFE module first applies a 3×3 depthwise separable convolution ($DWConv_{3\times3}$) to independently convolve each channel, followed by batch normalization (BN) to generate the transition feature $FL1$.

$$FL1 = BN(DWConv_{3\times3}FL) \tag{1}$$

Next, $FL1$ is first fused through a 1×1 convolution for cross-channel information integration and then processed with a nonlinear activation function Φ to maintain feature distribution characteristics, enhancing the network's ability to model complex features. A second 1×1 convolution achieves a nonlinear transformation of the feature channel dimensions while keeping the spatial dimensions unchanged. Ultimately, the local spatial features of the voxels are integrated through a residual connection ($\oplus$) with the original input *FL*, significantly improving the network's robustness to dimensional changes. The specific formula is as follows:

$$FL_{out} = Conv_{1\times1}(\Phi(Conv_{1\times1}(FL1))) \oplus FL \tag{2}$$

where FL_{out} is the output feature and Φ is the activation function. The VCFE module effectively captures the local spatial information of the input feature map and, by fusing it with the original input, reduces the impact of missing spatial information on detection results.

3. Results

3.1. Datasets

nuScenes [41] is an outdoor autopilot dataset that provides different annotations for various tasks. In this study, Mean Average Precision (mAP) and nuScenes Detection Score (NDS) are used as the main evaluation metrics.

Waymo-mini is a subset of the Waymo open dataset [42], which requires less computational resources than the full dataset and is more suitable for rapid model validation. Based on the proportion of "day", "night", and "dawn" in the original data, Waymo-mini maintains the consistency of scene distribution (as shown in Table 1). Figure 4 shows typical samples of different time periods in this dataset.

Table 1. The proportions of "day", "night", and "dawn" in the Waymo open and Waymo-mini datasets.

Datasets		Day	Night	Dawn	All
Waymo open dataset	Train	646	79	73	798
	Percentage of training	80.95%	9.90%	9.15%	100%
	Val	160	23	19	202
	Percentage of evaluations	79.21%	11.39%	9.41%	100%
Waymo-mini	Train	113	14	13	140
	Percentage of training	80.71%	10%	9.29%	100%
	Val	28	4	3	35
	Percentage of evaluations	80%	11.43%	8.57%	100%

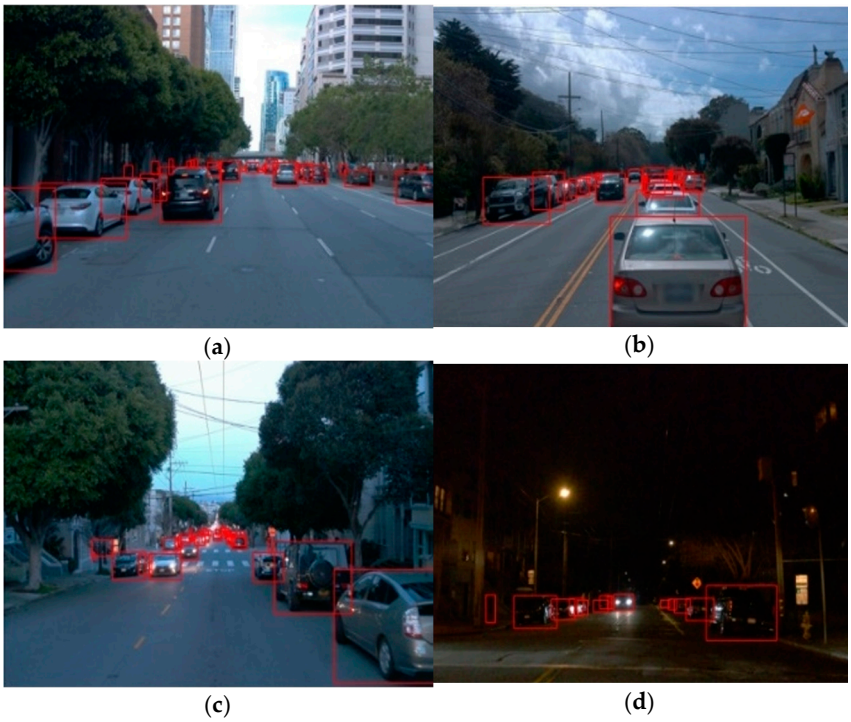


Figure 4. Number of objects in the Waymo-mini dataset at different time points. (**a**) and (**b**) correspond to "daytime". (**c**) corresponds to "dawn". (**d**) corresponds to "night". Targets to be detected are marked with red boxes.

Table 1 shows that the weights of "day", "night", and "dawn" scenes in Waymo-mini are similar to the original Waymo data. Figure 4 visualizes some samples from the Waymo-mini dataset, where each target to be detected is marked with a red box, and each image contains multiple targets to be detected. Like the original dataset, the evaluation protocol for the Waymo-mini dataset includes metrics such as average precision (AP) and average precision weighted by heading (APH). The detection results for each category were evaluated on two difficulty levels.

3.2. Experimental Details

In the nuScenes training scheme, the detection ranges along the X, Y, and Z axes were (−54, 54) m, (−54, 54) m, and (−5, 3) m, respectively, with voxel sizes of 0.3 m, 0.3 m, and 0.2 m, and all models were trained using the Adam optimizer, with a weight decay of 0.05 and a learning rate of 0.005, with a batch size of 3 and 20 epochs. For the Waymo-mini dataset, the detection ranges along the X, Y, and Z axes are (−74.88, 74.88) m, (−74.88, 74.88) m, and (−2, 4) m, and the voxel sizes are 0.32 m, 0.32 m, and 0.1875 m, respectively. The model was optimized using the Adam optimizer with a fixed weight decay set to 0.01

and a batch size set to 2 for 24 epochs. All experimental results were obtained on NVIDIA GeForce RTX 4090D GPUs and relied on OpenPCDet [43].

3.3. Experimental Results

Table 2 shows the detection performance of LDGF and mainstream methods on Waymo-mini. LDGF achieves leading accuracy on vehicle, pedestrian, and cyclist targets, and especially excels in small target (pedestrian and cyclist) detection. At Level 1, the average accuracies of pedestrian and cyclist reach 80.46% and 74.98%, respectively, which are 0.71% and 0.66% better than the existing optimal method, DSVT. The detection accuracies for vehicle targets reach 72.84% and 65.3% at Level 1 and Level 2, respectively, which are 0.66% and 0.71% higher than DSVT.

Table 2. The results on Waymo-mini have the optimal results of each column bolded.

Method	Vehicle				Pedestrian				Cyclist			
	Level1		Level2		Level1		Level2		Level1		Level2	
	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH
Pointpillar [15]	61.67	60.83	54.34	53.58	58.33	52.73	53.02	48.02	56.18	55.76	54.8	53.98
Second [16]	61.72	61.04	54.34	53.72	58.69	53.46	53.43	48.92	53.85	52.34	52.12	51.4
DSVT [23]	72.18	71.53	64.59	63.94	79.75	74.25	73.92	68.67	74.32	73.28	71.89	70.86
PVRCNN [24]	68.2	67.4	60.35	59.63	67.32	62.78	61.36	56.9	67.63	66.16	65.46	64.32
RDioU [44]	64.62	64	56.98	56.42	-	-	-	-	-	-	-	-
GCioU [45]	64.68	64.29	57.18	56.67	-	-	-	-	-	-	-	-
PartA2 [46]	66.72	66.09	58.93	58.36	65.67	60.4	59.72	54.28	64.08	63.83	62.02	61.62
M3DETR [47]	67.31	66.56	59.48	58.81	63.78	58.63	57.92	52.45	66.59	65.15	64.45	63.18
LDGF (ours)	**72.84**	**72.29**	**65.3**	**64.79**	**80.46**	**75.15**	**74.75**	**69.69**	**74.98**	**74.08**	**72.58**	**71.7**

Table 3 shows the detection times of different methods on the Waymo-mini dataset. Although the detection time required by LDGF is not the shortest, when combined with the accuracy of object detection, it demonstrates that LDGF has certain advantages.

Table 3. Detection time on the Waymo-mini dataset.

Methods	Time Indicator	
	Latency	FPS
PointPillars [15]	**41.12**	**24.32**
SECOND [16]	58.83	17
DSVT [23]	213	3.69
PV-RCNN [24]	408.83	2.45
RDIoU [44]	71.45	14
GCIoU [45]	64.93	15.4
PartA2 [46]	141.3	7.08
M3DETR [47]	895.15	1.07
LDGF (ours)	152.2	6.57

In Figure 5a–c, all of the images demonstrate pedestrians that can be detected normally, even when partially occluded or disproportionately small. Figure 5d, however, shows detection failure due to the pedestrian being too small in the frame. The green box in Figure 5d marks the actual target location, while the red box indicates the detected bounding box.

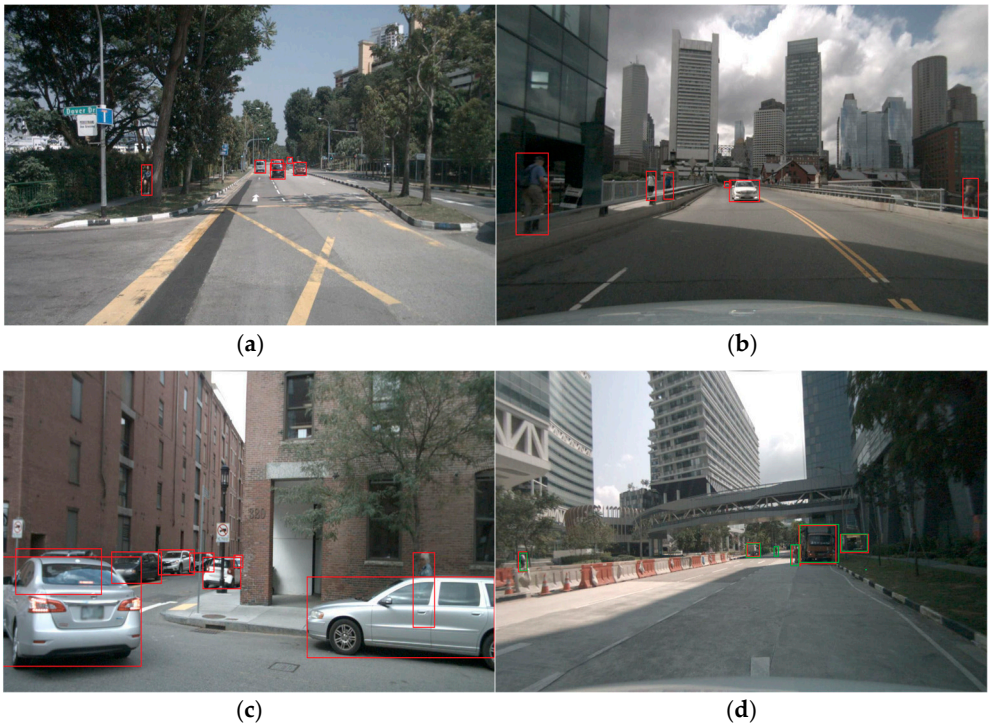


Figure 5. Detection images for small objects or pedestrians being blocked: images (**a**–**c**) indicate successful detection; image (**d**) indicates detection failure.

As shown in Table 4, on the nuScenes dataset, LDGF slightly outperforms the current optimal method, DSVT, in both NDS and mAP metrics. In particular, LDGF achieves 88.1% and 77.7% detection accuracy in typical traffic scene targets, such as car and bus, respectively, indicating that the proposed global modeling and channel enhancement strategy has good generalization ability. For motor, bicycle, and barrier targets, due to their low percentage in the nuScenes dataset, LDGF does not have sufficient learning ability for these targets, leading to detection results that are not the best. However, Table 5 further presents a performance comparison between LDGF and multi-modal 3D object detection methods that fuse camera and radar data. The results demonstrate that LDGF's mAP (66.9) and NDS (71.4) outperform most multi-modal methods, including CRN [48] (mAP 57.5, NDS 62.4), CRAFT [49] (mAP 41.1, NDS 52.3), RCBEVDet [50] (mAP 55.0, NDS 63.9), and RCBEVDet++ [51] (ResNet50) (mAP 51.9, NDS 60.4). Only RCBEVDet++ (ViT-Large) (mAP 67.3, NDS 72.7) slightly surpasses LDGF. Nevertheless, as a single-modality method, LDGF avoids the complexities of sensor calibration and sensitivity to radar sparsity inherent in multi-modal approaches, while maintaining lower computational complexity and efficient small object detection. This indicates that, except when using the ViT-Large backbone, camera radar-based multi-modal methods struggle to outperform LDGF. Overall, the LDGF constructed in this paper achieves robust detection results on the nuScenes dataset, highlighting its competitiveness and simplicity in autonomous driving perception tasks.

In Waymo-mini and nuScenes, two datasets with high real-world perceptual complexity, the LDGF achieves stable detection results under multiple targets, different weather, illumination, and density environments. In particular, the detection accuracy for small targets shows certain advantages, which, side by side, confirms the modeling robustness of the structure in this paper when facing perceptual uncertainties.

Table 4. The results of the nuScenes dataset show that the optimal results of each column are bolded.

	NDS	mAP	Car	Truck	Bus	Trailer	C.V	Ped.	Motor	Bicycle	T.C.	Barrier
3DSSD [10]	56.4	42.7	81.2	47.2	61.4	30.5	12.6	70.2	36.0	8.6	31.1	47.9
Pointpillar [15]	44.9	29.5	70.5	25.0	34.4	20.0	4.5	59.9	16.7	1.6	29.6	33.2
Second [16]	-	27.1	75.5	21.9	29.0	13.0	0.4	59.9	16.9	0	22.5	32.2
DSVT [23]	71.1	66.4	87.4	62.6	75.9	42.1	25.3	88.2	74.8	58.7	77.8	**70.9**
SASA [52]	61.0	45.0	76.8	45.0	66.2	36.5	16.1	69.1	39.6	16.9	29.9	53.6
TransFusion-L [53]	70.1	65.5	86.9	60.8	73.1	43.4	25.2	**87.5**	72.9	57.3	77.2	70.3
FCOS-LiDAR [54]	57.1	63.2	82.1	52.3	65.2	33.6	18.3	84.1	58.5	35.3	73.4	67.9
PVT-SSD [28]	65.0	53.6	79.4	43.48	62.1	34.2	21.7	79.8	53.4	38.2	56.6	67.1
LDGF (ours)	**71.4**	**66.9**	**88.1**	**63.8**	**77.7**	**45.0**	**28.6**	88.2	**74.1**	**58.2**	**78.7**	66.6

Table 5. Comparison of LDGF with multi-modal methods on nuScenes.

Method	Modalities	Drawbacks	mAP	NDS
CRN [48]	Camera + Radar	Complex sensor calibration, moderate computational overhead	57.5	62.4
CRAFT [49]	Camera + Radar	High training data demand, sensitive to radar sparsity	41.1	52.3
RCBEVDet [50]	Camera + Radar	Real-time performance is hardware-dependent, lower depth precision vs. LiDAR	55.0	63.9
RCBEVDet++ [51] (ResNet50)	Camera + Radar	Increased computational complexity, sensitive to radar point sparsity	51.9	60.4
RCBEVDet++ (ViT-Large)	Camera + Radar		**67.3**	**72.7**
LDGF (ours)	LiDAR (only)	Limited to single modality, no multi-modal gains	66.9	71.4

3.4. Ablation Experiments

To verify the effectiveness of the PFEB and VCFE modules, Table 6 demonstrates the detection effectiveness of each module on Waymo-mini for car and pedestrian targets. After adding PFEB, the detection accuracy of the network for the car class at Level 1 and Level 2 is improved by 0.15% and 0.2%, respectively; the detection accuracy of the network for the pedestrian class at both levels is improved by 0.41% and 0.45%, respectively, which indicates that PFEB effectively extends the network sensing field. VCFE is a further optimization of the performance of PFEB, and after joining its network, it improves the detection accuracies for car class targets by 0.51% and 0.47% and for pedestrian class targets by 0.3% and 0.38%, respectively, which verifies the importance of local spatial complementation for point cloud target detection, especially for small targets.

As shown in Table 7, adopting a unidirectional sequence (Sq+) yields comparable or slightly improved performance over the baseline across both car and pedestrian categories. In contrast, the bidirectional sequence (Sq+ + Sq−) consistently achieves further gains, particularly on pedestrian detection, where Level 1 AP improves from 79.75 to 80.16, and Level 2 AP improves from 73.92 to 74.37. These results validate that the bidirectional traversal effectively alleviates directional bias in voxel serialization and enhances global feature completeness, leading to more robust detection of small and occluded objects.

Table 6. The effect of each module of LDGF, and the optimal result of each column is bolded.

Ablation	Car				Pedestrian			
	Level 1		Level 2		Level 1		Level 2	
	AP	APH	AP	APH	AP	APH	AP	APH
Baseline	72.18	71.53	64.59	63.94	79.75	74.25	73.92	68.67
+PFEB	72.33	71.82	64.79	64.32	80.16	74.47	74.37	68.97
+VCFE	**72.84**	**72.29**	**65.53**	**64.79**	**80.46**	**75.15**	**74.75**	**69.69**

Table 7. Ablation study of unidirectional (Sq+) and bidirectional (Sq+ + Sq−) sequences on car and pedestrian detection.

Ablation	Car				Pedestrian			
	Level 1		Level 2		Level 1		Level 2	
	AP	APH	AP	APH	AP	APH	AP	APH
Baseline	72.18	71.53	64.59	63.94	79.75	74.25	73.92	68.67
Sq+	72.21	71.68	65.33	64.23	79.81	74.27	74.15	68.58
Sq+ Sq−	**72.33**	**71.82**	**64.79**	**64.32**	**80.16**	**74.47**	**74.37**	**68.97**

The VMB is merely a mapping module designed to transform point cloud data into long data sequences for subsequent GFEB extraction of global features. Essentially, the VMB and GFEB together form PFEB, which cannot be separated and must operate as an integrated unit.

Table 8 presents the detection performance of the network under various channel division strategies of the VCFE (Voxel-based Channel Feature Enhancement) module, evaluated using a laser point cloud with a data channel dimension of 128. Due to this dimension, the channel size for division must be an exponential multiple of 2 (i.e., 1, 2, 4, or 8). Additionally, the PFEB (Point Feature Enhancement Block) module is incorporated when the channel size is set to 1, serving as the baseline configuration for comparison. The table reports the average precision (AP) and Average Precision with Heading (APH) metrics for two object classes—car and pedestrian—across two difficulty levels (Level 1 and Level 2).

Table 8. The detection results under different channel divisions of VCFE, and the optimal results of each column are bolded.

Channel Size	Car				Pedestrian			
	Level1		Level2		Level1		Level2	
	AP	APH	AP	APH	AP	APH	AP	APH
1	72.33	71.82	64.79	64.32	80.16	74.47	74.37	68.97
2	72.67	72.15	65.09	64.61	80.20	74.55	74.33	68.97
4	**72.84**	**72.29**	**65.53**	**64.79**	**80.46**	**75.15**	**74.75**	**69.69**
8	72.58	72.06	64.99	64.5	79.72	74.19	73.93	68.68

From the results in Table 8, it is evident that the detection performance is significantly influenced by the choice of channel size. The optimal performance across all metrics is achieved when the channel size is set to 4, as indicated by the bolded values in the table. Specifically, for the car class, the AP and APH scores at Level 1 are 72.84 and 72.29, respectively, while at Level 2, they are 65.53 and 64.79. For the pedestrian class, the AP and APH scores at Level 1 are 80.46 and 75.15, and at Level 2, they are 74.75 and 69.69,

respectively. These results consistently outperform other channel sizes, demonstrating that a channel size of 4 strikes an optimal balance in capturing and modeling spatial information for enhanced detection accuracy.

When the channel size is increased to 8, there is a slight degradation in performance compared to the channel size of 4. For instance, the AP for the car class at Level 1 drops to 72.58, and for the pedestrian class at Level 1, it decreases to 79.72. Despite this decline, the performance with a channel size of 8 still surpasses the baseline configuration (channel size of 1, with the PFEB module), where the AP for car at Level 1 is 72.33, and for pedestrian at Level 1, it is 80.16. This comparison highlights the effectiveness of the VCFE module in improving the modeling of spatial information, even when the channel size is not optimal.

The superior performance at a channel size of 4 can be attributed to its ability to effectively balance the granularity of feature extraction and the computational complexity of the network. Smaller channel sizes (e.g., 1 or 2) may lead to overly coarse feature representations, while a larger channel size (e.g., 8) may introduce redundancy or dilute the feature focus, resulting in suboptimal detection performance. Consequently, the LDGF (Laser-based Detection and Grounding Framework) adopts a channel size of 4 for the VCFE module to maximize detection accuracy while maintaining computational efficiency.

In conclusion, Table 8 underscores the critical role of the VCFE module in enhancing the network's ability to model spatial information from laser point cloud data. The choice of a channel size of 4 optimizes detection performance across both car and pedestrian classes, validating the design of the VCFE module and its integration into the LDGF framework.

4. Conclusions

In this paper, we introduce a novel LiDAR point cloud object detection algorithm, termed the Laser-based Detection and Grounding Framework (LDGF), which innovatively incorporates global feature extraction to address the limitations of traditional local convolution-based methods. The LDGF framework leverages a Voxel Mapping Block (VMB) to transform three-dimensional point cloud data into a one-dimensional long sequence, enabling efficient processing and global feature modeling. Subsequently, the Global Feature Extraction Block (GFEB), inspired by the selective state space model (SSM) and the Mamba architecture, extracts comprehensive global contextual features with approximately linear computational complexity. This approach significantly enhances the network's receptive field, allowing it to capture long-range dependencies and improve detection performance, particularly for small targets such as pedestrians and cyclists in complex autonomous driving scenarios.

To mitigate the loss of local spatial information caused by the one-dimensional transformation, we propose the Voxel Channel Feature Extraction (VCFE) module. This module divides and correlates point cloud features across different channels, effectively capturing local spatial information and compensating for the spatial degradation introduced during data transformation. The VCFE module's ability to balance local and global feature extraction ensures robust detection performance, as demonstrated by the experimental results on the Waymo-mini and nuScenes datasets. Specifically, LDGF achieves superior detection accuracy compared to state-of-the-art methods, with improvements of 0.64%, 0.71%, and 0.66% in average precision for vehicle, pedestrian, and cyclist targets, respectively, on the Waymo-mini dataset. On the nuScenes dataset, LDGF improves car detection accuracy by 0.7%, with NDS and mAP increasing by 0.3% and 0.5%, respectively. The ablation studies further validate the contributions of the PFEB and VCFE modules, with the optimal channel size of 4 in the VCFE module striking a balance between feature granularity and computational efficiency, leading to enhanced detection performance.

The LDGF framework demonstrates remarkable generalization and robustness across diverse environmental conditions, including varying weather, illumination, and point cloud densities. Its standout performance in small target detection underscores its potential to address critical challenges in autonomous driving perception systems, where detecting smaller objects like pedestrians and cyclists is often hindered by limited receptive fields in traditional methods. By integrating global feature modeling with local spatial enhancement, LDGF not only pushes the boundaries of point cloud object detection but also sets a foundation for future research in scalable and efficient 3D perception systems.

Looking ahead, the proposed approach opens several avenues for further exploration. Potential improvements include optimizing the computational efficiency of the VMB and GFEB modules for real-time applications, exploring adaptive channel division strategies in the VCFE module to handle varying point cloud densities, and extending the framework to multi-modal perception systems that integrate LiDAR with other sensor data, such as cameras or radar. Additionally, further investigations could focus on enhancing the robustness of LDGF in extreme scenarios, such as dense urban environments or adverse weather conditions, to ensure its reliability in real-world autonomous driving applications. In conclusion, LDGF represents a significant advancement in LiDAR-based object detection, offering a robust, efficient, and scalable solution that enhances the safety and reliability of autonomous driving systems.

Author Contributions: Conceptualization, X.Z. and Y.L.; methodology, Y.L. and F.T.; software, F.T.; validation, J.S. and F.T.; formal analysis, Y.L.; investigation, X.Z. and J.S.; resources, Y.L.; data curation, Y.L. and J.S.; writing—original draft preparation, X.Z.; writing—review and editing, F.T.; visualization, F.T.; supervision, X.Z.; project administration, Y.L. and X.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Henan Science and Technology R&D Program Joint Fund Project (Young Scientists), grant number 225200810098, and the Key R&D and Promotion Projects of Henan Province (Science and Technology Research), grant number 242102211008.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available at https://www.nuscenes.org/nuscenes (accessed on 23 September 2025) and https://waymo.com/open (accessed on23 September 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Y.; Wen, J.; Gong, R.; Ren, B.; Li, W.; Cheng, C.; Liu, H.; Sebe, N. PVAFN: Point-Voxel Attention Fusion Network with Multi-Pooling Enhancing for 3D Object Detection. *Expert Syst. Appl.* **2025**, *281*, 127608. [CrossRef]
2. Zheng, Q.; Wu, S.; Wei, J. VoxT-GNN: A 3D Object Detection Approach from Point Cloud Based on Voxel-Level Transformer and Graph Neural Network. *Inf. Process. Manag.* **2025**, *62*, 104155. [CrossRef]
3. Zhang, C.; Wang, H.; Cai, Y.; Chen, L.; Li, Y. TransFusion: Transformer-Based Multi-Modal Fusion for 3D Object Detection in Foggy Weather Based on Spatial Vision Transformer. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 10652–10666. [CrossRef]
4. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [CrossRef]
5. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5099–5108.
6. Choy, C.; Gwak, J.Y.; Savarese, S. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3075–3084.
7. Deng, J.; Zhang, S.; Dayoub, F.; Ouyang, W.; Zhang, Y.; Reid, I. PoIFusion: Multi-Modal 3D Object Detection via Fusion at Points of Interest. *arXiv* **2024**, arXiv:arXiv:2403.09212.
8. Zhang, Y.; Wang, Y.; Cui, Y.; Chau, L.-P. 3DGeoDet: General-purpose geometry-aware image-based 3D object detection. *arXiv* **2025**, arXiv:arXiv:2506.09541.

9. Hamilton, J.D. State-space models. *Handb. Econom.* **1994**, *4*, 3039–3080.
10. Yang, Z.; Sun, Y.; Liu, S.; Jia, J. 3dssd: Point-based 3d single stage object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11040–11048.
11. Zhang, Y.; Hu, Q.; Xu, G.; Ma, Y.; Wan, J.; Guo, Y. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18953–18962.
12. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 770–779.
13. Graham, B.; Engelcke, M.; Van Der Maaten, L. 3D semantic segmentation with submanifold sparse convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9224–9232.
14. Graham, B.; Van der Maaten, L. Submanifold sparse convolutional networks. *arXiv* **2017**, arXiv:1706.01307.
15. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 12697–12705.
16. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. *Sensors* **2018**, *18*, 3337. [CrossRef] [PubMed]
17. Zheng, W.; Tang, W.; Chen, S.; Jiang, L.; Fu, C.-W. Cia-ssd: Confident iou-aware single-stage object detector from point cloud. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 3555–3562. [CrossRef]
18. Wang, H.; Chen, Z.; Cai, Y.; Chen, L.; Li, Y.; Sotelo, M.A.; Li, Z. Voxel-RCNN-Complex: An effective 3D point cloud object detector for complex traffic conditions. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [CrossRef]
19. Liu, W.; Zhu, D.; Luo, H.; Li, Y. 3D Object Detection in LiDAR Point Clouds Fusing Point Attention Mechanism. *Acta Photonica Sin.* **2023**, *52*, 221–231.
20. Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; Jia, J. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21674–21683.
21. An, P.; Duan, Y.; Huang, Y.; Ma, J.; Chen, Y.; Wang, L.; Yang, Y.; Liu, Q. SP-Det: Leveraging Saliency Prediction for Voxel-based 3D Object Detection in Sparse Point Cloud. *IEEE Trans. Multimed.* **2024**, *26*, 2795–2808. [CrossRef]
22. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-based 3d object detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11784–11793.
23. Wang, H.; Shi, C.; Shi, S.; Lei, M.; Wang, S.; He, D.; Schiele, B.; Wang, L. Dsvt: Dynamic sparse voxel transformer with rotated sets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13520–13529.
24. 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, Seattle, WA, USA, 13–19 June 2020; pp. 10529–10538.
25. Shi, S.; Jiang, L.; Deng, J.; Wang, Z.; Guo, C.; Shi, J.; Wang, X.; Li, H. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection. *Int. J. Comput. Vis.* **2023**, *131*, 531–551. [CrossRef]
26. Sheng, H.; Cai, S.; Liu, Y.; Deng, B.; Huang, J.; Hua, X.S.; Zhao, M.J. Improving 3d object detection with channel-wise transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2743–2752.
27. Liu, H.; Dong, Z.; Tian, S. Object Detection Network Fusing Point Cloud and Voxel Information. *Comput. Eng. Des.* **2024**, *45*, 2771–2778. [CrossRef]
28. Yang, H.; Wang, W.; Chen, M.; Lin, B.; He, T.; Chen, H.; He, X.; Ouyang, W. Pvt-ssd: Single-stage 3d object detector with point-voxel transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13476–13487.
29. Deng, P.; Zhou, L.; Chen, J. PVC-SSD: Point-Voxel Dual-Channel Fusion with Cascade Point Estimation for Anchor-Free Single-Stage 3D Object Detection. *IEEE Sens. J.* **2024**, *24*, 14894–14904. [CrossRef]
30. Tian, F.; Wu, B.; Zeng, H.; Zhang, M.; Hu, Y.; Xie, Y.; Wen, C.; Wang, Z.; Qin, X.; Han, W.; et al. A shape-attention Pivot-Net for identifying central pivot irrigation systems from satellite images using a cloud computing platform: An application in the contiguous US. *GIScience Remote Sens.* **2023**, *60*, 2165256. [CrossRef]
31. Liu, M.; Wang, W.; Zhao, W. PVA-GCN: Point-voxel absorbing graph convolutional network for 3D human pose estimation from monocular video. *Signal Image Video Process.* **2024**, *18*, 3627–3641. [CrossRef]
32. Xu, W.; Fu, T.; Cao, J.; Zhao, X.; Xu, X.; Cao, X.; Zhang, X. Mutual information-driven self-supervised point cloud pre-training. *Knowl.-Based Syst.* **2025**, *307*, 112741. [CrossRef]
33. Spatial Sparse Convolution Library. 2022. Available online: https://github.com/traveller59/spconv (accessed on 23 September 2025).
34. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv* **2021**, arXiv:2111.00396.

35. Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; Ré, C. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 572–585.
36. Smith, J.T.H.; Warrington, A.; Linderman, S.W. Simplified state space layers for sequence modeling. *arXiv* **2022**, arXiv:2208.04933.
37. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417. [CrossRef]
38. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063.
39. Yu, W.; Wang, X. Mambaout: Do we really need mamba for vision? *arXiv* **2024**, arXiv:2405.07992. [CrossRef]
40. Zhou, Y.; Sun, P.; Zhang, Y.; Anguelov, D.; Gao, J.; Guo, J.; Ngiam, J.; Vasudevan, V. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In Proceedings of the Conference on Robot Learning, Virtual, 16–18 November 2020; pp. 923–932.
41. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
42. 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, Seattle, WA, USA, 13–19 June 2020; pp. 2446–2454.
43. Liu, Y.X.; Pan, X.; Zhu, J. 3D Pedestrian Detection Based on Pointpillar: SelfAttention-pointpillar. In Proceedings of the 2024 9th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 21–23 November 2024; Volume 9, pp. 1–10.
44. Sheng, H.; Cai, S.; Zhao, N.; Deng, B.; Huang, J.; Hua, X.-S.; Zhao, M.-J.; Lee, G.H. Rethinking IoU-based optimization for single-stage 3D object detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 544–561.
45. Ming, Q.; Miao, L.; Ma, Z.; Zhao, L.; Zhou, Z.; Huang, X.; Chen, Y.; Guo, Y. Deep dive into gradients: Better optimization for 3D object detection with gradient-corrected IoU supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5136–5145.
46. Shi, S.; Wang, Z.; Shi, J.; Wang, X.; Li, H. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2647–2664. [CrossRef]
47. Guan, T.; Wang, J.; Lan, S.; Chandra, R.; Wu, Z.; Davis, L.; Manocha, D. M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 772–782.
48. Kim, Y.; Shin, J.; Kim, S.; Lee, I.-J.; Choi, J.W.; Kum, D. RN: Camera Radar Net for Accurate, Robust, Efficient 3D Perception. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023), Paris, France, 2–6 October 2023; pp. 17569–17580.
49. Kim, Y.; Kim, S.; Choi, J.W.; Kum, D. Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 1160–1168. [CrossRef]
50. Lin, Z.; Liu, Z.; Xia, Z.; Wang, X.; Wang, Y.; Qi, S.; Dong, Y.; Dong, N.; Zhang, L.; Zhu, C. Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 14928–14937.
51. Lin, Z.; Liu, Z.; Wang, Y.; Zhang, L.; Zhu, C. RCBEVDet++: Toward high-accuracy radar-camera fusion 3D perception network. *arXiv* **2024**, arXiv:2409.04979.
52. Chen, C.; Chen, Z.; Zhang, J.; Tao, D. Sasas: Semantics-augmented set abstraction for point-based 3d object detection. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 221–229. [CrossRef]
53. Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; Tai, C.-L. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Orleans, LA, USA, 18–24 June 2022; pp. 1090–1099.
54. Tian, Z.; Chu, X.; Wang, X.; Wei, X.; Shen, C. Fully convolutional one-stage 3d object detection on lidar range images. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 34899–34911.

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.