Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 20, No. 1, 2026

A Sparse Point Cloud 3D Object Detection Method Based on Size-Aware Attention Mechanism and BEV Backbone Network

Ying He, Chenglong Wang, Chuanjiang Yan

School of Computer Science & Engineering, Chongqing University of Technology, Chongqing 400054, China

Abstract: Addressing the challenge of effectively detecting distant and small-scale objects in sparse point cloud scenarios, this paper proposes an improved 3D object detection method based on a size-aware attention mechanism and a BEV backbone network, achieving performance enhancements within the PV-RCNN framework. First, the Size-Aware Dual-Branch Attention mechanism is introduced during the keypoint feature fusion stage. By estimating the voxel density in keypoint neighborhoods, it dynamically determines target scales. Channel attention and spatial attention are applied differentially to enhance keypoint features, thereby improving the models ability to represent multi -scale features. Second, the BaseBEV backbone network is designed and integrated into the BEV feature extraction stage, replacing the original SECOND and SECONDFPN structures to enhance multi-scale feature modeling capabilities in the BEV space. Experimental validation on the KITTI 3D object detection dataset demonstrates that the proposed method achieves consistent improvements in both BEV detection and 3D detection tasks. Ablation studies further validate the effectiveness of each improved module, confirming that our approach effectively enhances 3D object detection performance in sparse point cloud environments.

Keywords: 3D object detection; PV-RCNN; Size-aware attention; BEV backbone network.

1. Introduction

With the rapid advancement of autonomous driving, intelligent robotics, and drone technologies, 3D object detection has emerged as a critical research focus in computer vision and robotic perception [1]. In autonomous driving scenarios, vehicles rely on 3D object detection technology to accurately and in real-time obtain the three-dimensional position, size, and category information of surrounding objects, enabling safe and reliable decision-making. The accuracy and real-time performance of this technology directly determine the safety and reliability of autonomous driving [2].

However, point cloud data inherently exhibits characteristics such as high sparsity, uneven distribution, and severe occlusion, which are particularly pronounced in long-range and small-object detection tasks. These traits make feature representation challenging and compromise classification and localization accuracy [3]. To address these challenges, researchers have proposed numerous deep learning-based 3D detection methods. Based on how point cloud data is processed, these methods can be broadly categorized into point-based methods, voxel-based methods, and point-voxel fusion methods. Point-based methods perform feature learning directly on raw point clouds. For instance, PointNet/PointNet++ achieve end-to-end modeling through point set feature encoding [4,5], which further evolved into two-stage point cloud detection frameworks like PointRCNN [6]. Voxel-based methods map point clouds onto regular grids via voxelization to adapt them for convolutional networks. Representative works include VoxelNet and SECOND, which offer high computational efficiency [7,8]. Additionally, PointPillars employs pillar representations and performs 2D convolutional modeling in BEV space, achieving a good balance between efficiency and accuracy [9]. However, the quantization errors introduced by voxelization may weaken the representation of local geometric details, particularly in scenarios involving small objects and sparse regions.

To balance the global modeling capabilities of voxel features with the fine-grained geometric representation of point features, point-voxel fusion methods have emerged as a significant research focus in recent years. Among these, PV-RCNN achieves outstanding performance on the KITTI dataset by introducing the Voxel-to-Keypoint Feature Abstraction (VSA) mechanism to fuse multi-scale voxel features with point features at keypoints, combined with RoI-grid pooling for precise bounding box regression [10]. Despite PV-RCNNs strong detection performance, its fixed BEV feature backbone structure limits spatial context modeling in complex scenes. Additionally, its uniform keypoint feature enhancement strategy struggles to adapt to varying point cloud densities and object scales, particularly for small targets like pedestrians and cyclists. On the other hand, attention mechanisms widely adopted in 2D vision tasks (e.g., SENet, CBAM) can enhance feature representation across both channel and spatial dimensions [11,12], offering valuable insights for keypoint feature modeling.

Based on the above analysis, this paper proposes SAPV-RCNN, a sparse point cloud 3D object detection method enhanced with size-aware attention and a BEV backbone network. While retaining the two-stage detection workflow of PV-RCNN, improvements are made in two key aspects: (1) Introducing the BaseBEV backbone network to replace the original SECOND and feature pyramid structures, enhancing multi-scale feature expression in BEV space; (2) Integrating a size-aware dual-branch attention mechanism after keypoint feature fusion. This mechanism adaptively determines target scale based on voxel density around keypoints, applying channel attention and spatial attention respectively to differentially enhance keypoint features. This approach

improves detection performance for small objects and sparse point cloud regions. Experimental validation on the KITTI dataset demonstrates that the proposed method achieves detection performance superior to the PV-RCNN baseline in both BEV and 3D detection tasks, confirming the effectiveness of the improvement strategy.

2. Related Work

In recent years, significant progress has been made in three-dimensional object detection methods based on LiDAR point clouds. Due to the sparse, irregular, and unevenly distributed nature of point cloud data, maintaining geometric detail representation while ensuring computational efficiency has become one of the core challenges in this field. In alignment with the research objectives of this paper, this chapter focuses on reviewing two closely related research directions: detection frameworks based on voxel/BEV representations and two-stage detection methods that fuse points and voxels. Building upon this review, we analyze the limitations of existing approaches, thereby providing the research context and comparative basis for the subsequent SAPV-RCNN method.

Bird's Eye View (BEV) representations map point clouds onto regular two-dimensional grids, enabling efficient feature modeling by convolutional neural networks on the BEV plane. Consequently, BEV representations are widely adopted for 3D object detection in autonomous driving. Early work like PIXOR directly predicts bounding boxes in BEV space, achieving high detection efficiency [13]. Subsequently, VoxelNet introduced the Voxel Feature Encoding (VFE) mechanism combined with 3D convolutions to achieve end-to-end detection, laying the foundation for voxel-based detection frameworks [7]. SECOND introduced sparse convolutional structures, significantly improving feature encoding efficiency and becoming a classic representative of voxel-based methods [8]. Additionally, PointPillars partitions point clouds into pillar structures and models them via 2D convolutions on BEV planes, achieving a favorable balance between accuracy and real-time performance [9].

In recent years, the BEV-based keypoint detection paradigm has further evolved. For instance, CenterPoint predicts center points in BEV space and regresses 3D bounding boxes, demonstrating outstanding performance across multiple benchmark datasets [14]. Simultaneously, BEV representations have increasingly integrated with multimodal information fusion. BEVFusion demonstrates the frameworks scalability by enhancing detection robustness through multi-sensor feature fusion within a unified BEV space [15].

Overall, voxel/BEV methods offer computational efficiency and parallelizability advantages, yet their detection performance heavily relies on the feature representation capabilities of the BEV backbone network. Particularly in complex road scenes, where multi-scale object distributions vary significantly, insufficient spatial context and multi-scale information fusion capabilities in the BEV backbone network may compromise candidate box generation quality and overall detection accuracy. Therefore, research focused on optimizing BEV backbone network architecture and enhancing multi-scale feature representation capabilities holds significant importance.

Pure voxel/BEV representations may introduce quantization errors, weakening the expression of fine-grained geometric information in point clouds. To address this, point-voxel fusion methods—which combine the global expressive power of voxel features with the geometric detail advantages of point features—have emerged as a key research direction in 3D object detection in recent years. Part-A2 adopts a two-stage detection framework and introduces a local feature enhancement strategy, improving detection accuracy by refining the modeling of point cloud features within candidate bounding boxes [16]. Building upon this, PV-RCNN proposes the Voxel-to-Keypoint Feature Abstraction (VSA) mechanism, which fuses multi-scale voxel semantic features with point features at keypoints. It further aggregates features within candidate boxes via RoI-grid pooling to achieve high-precision 3D object detection [10].

To further enhance the feature representation capability and efficiency of the fusion framework, PV-RCNN++ optimizes the network architecture and keypoint feature abstraction strategy, improving performance and speed while retaining the advantages of two-stage detection [17]. Additionally, some studies have explored end-to-end or single-stage point cloud detection architectures. For instance, 3DSSD proposed a keypoint detection paradigm independent of candidate boxes, achieving end-to-end detection through sampling and feature enhancement [18]. These works indicate that keypoint feature representation and multi-scale information fusion remain critical factors influencing point cloud detection performance.

Overall, point-voxel fusion methods strike a reasonable balance between accuracy and efficiency, though room for improvement remains in complex scenarios. The BEV backbone structure determines candidate bounding box generation quality, while fixed backbone networks exhibit limited multiscale context modeling capabilities. Simultaneously, keypoint feature representation directly impacts the precision of candidate box regression, posing challenges in sparse point clouds and small-object scenarios. Therefore, improvements in both BEV feature modeling and keypoint feature enhancement are crucial. Based on this, we propose the SAPV-RCNN framework. It enhances BEV multi-scale feature representation by introducing the BaseBEV backbone network and refines keypoint feature modeling strategies to boost 3D object detection performance on sparse point clouds.

3. Method

3.1. Review of PV-RCNN Baseline Methods

PV-RCNN is a representative point-voxel fusion 3D object detection method. Its core idea is to introduce point-level features to enhance the representation of local geometric details while preserving the efficient global modeling capability of voxel representations.

In its overall workflow, PV-RCNN first voxelizes the input point cloud and encodes voxel features using a sparse convolutional network to generate high-level semantic feature maps in a birds -eye view (BEV). Based on these BEV features, the network generates initial 3D candidate boxes via a Region Proposal Network (RPN).

Following candidate box generation, PV-RCNN incorporates a Volumetric Sampling Module (VSA)—a key design distinguishing it from traditional voxel-based detection methods. Guided by candidate boxes, this module samples a set of representative keypoints from the original point cloud and fuses information from multi-scale voxel features and point features at these keypoint locations.

Through this approach, PV-RCNN simultaneously preserves the semantic information of voxel features and the geometric details of point-level features at keypoints, thereby mitigating the shortcomings of pure voxel representations in modeling local structures.

During the fine-grained regression stage, PV-RCNN further constructs a regular RoI-grid within candidate boxes and aggregates keypoint features to form fixed-dimensional RoI representations. Subsequently, the RCNN detection head performs fine-grained regression and classification predictions for the center position, size, and orientation of candidate boxes. This design enables the network to fully leverage keypoint features within local regions for more accurate candidate box refinement, significantly enhancing 3D detection accuracy.

Through the synergistic design of these three key stages, PV-RCNN achieves effective fusion of voxel features and point features while maintaining computational efficiency, delivering outstanding performance on multiple 3D object detection benchmarks.

3.2. SAPV-RCNN Overall Network Architecture

Although PV-RCNN achieves promising results in 3D object detection through point-voxel fusion, it still faces limitations in complex real-world road scenarios. On one hand, its fixed BEV backbone architecture has limited modeling capabilities for spatial structural information in complex scenes; On the other hand, during keypoint feature modeling, objects of varying scales share a uniform feature enhancement approach, making it difficult to effectively adapt to scenarios with significant object size variations and highly uneven point cloud distributions. Particularly, there remains room for improvement in detecting small-sized and distant objects.

To address these issues, this paper proposes a sparse point cloud 3D object detection framework (SAPV-RCNN) based on a size-aware attention mechanism and a BEV backbone network. While preserving the overall detection workflow and two-stage detection paradigm of PV-RCNN, this framework enhances the original network architecture by improving both BEV feature modeling and keypoint feature augmentation. This enhances the models 3D object detection performance in complex scenes. The overall framework is illustrated in Figure 1.

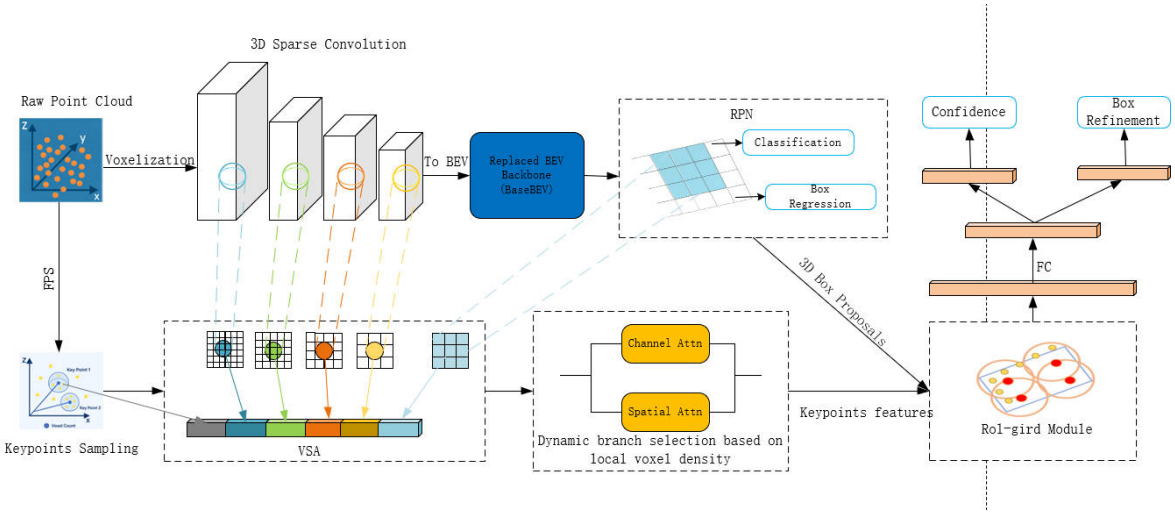

Figure 1. Schematic diagram of the overall SAPV-RCNN network architecture.

In the overall workflow, SAPV-RCNN still uses raw point clouds as input. It first performs voxelization on the point clouds and encodes the voxel features through a 3D sparse convolutional network. Subsequently, the encoded voxel features are projected into birds-eye view (BEV) space for subsequent bounding box candidate generation and global modeling. At this stage, the BaseBEV backbone network replaces the original backbone structure to enhance the expressive power of BEV features in spatial structure and semantic information. Based on the enhanced BEV features, the RPN generates initial 3D candidate boxes, providing reliable candidate regions for subsequent fine-grained detection.

SAPV-RCNN inherits PV-RCNNs keypoint detection strategy, sampling a set of keypoints from the original point cloud guided by candidate boxes. The VSA module then fuses multi-scale voxel features with point features to obtain keypoint feature representations. To enhance the adaptability of keypoint features to objects of varying scales, this paper introduces a size-aware dual-branch attention mechanism (SAD-Attention) after the VSA module. This module adaptively models target scales based on the voxel distribution features surrounding keypoints. It employs both channel attention and spatial attention to enhance keypoint features, thereby improving the models feature discrimination capability for small-sized objects and sparse point cloud regions.

Finally, SAPV-RCNN aggregates key point features within candidate boxes through the RoI-grid feature aggregation module. It then employs the RCNN detection head for precise regression and classification predictions of candidate box positions, sizes, and orientations, ultimately outputting 3D object detection results.

3.3. Design of a BEV Backbone Network Based on BaseBEV

In PV-RCNN, the BEV backbone network is responsible for modeling birds-eye-view features derived from voxel feature projections, providing high-quality spatial feature representations for the subsequent Region Proposal Network (RPN). The original PV-RCNN employs SECOND and its feature pyramid structure (SECONDFPN) as the BEV

backbone, primarily extracting multi-scale information through stepwise downsampling and simple feature fusion. However, this architecture exhibits relatively limited multiscale feature interactions within the BEV space, leaving room for improvement in modeling spatial contextual information for complex scenes. To address this issue, this paper introduces BaseBEV as a new BEV backbone network. While preserving the overall detection workflow, it implements structural improvements to the BEV feature modeling approach, as illustrated in Figure 2.

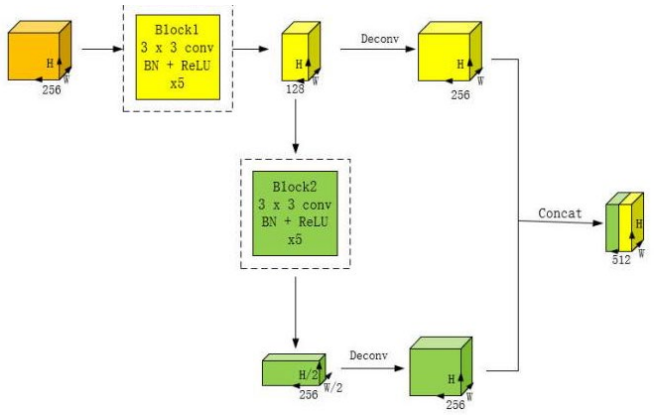

Figure 2. Schematic diagram of the BaseBEV backbone network architecture.

Let the BEV input feature $X_0 \in \mathbb{R}^{B\times C_0\times H\times W}$, be obtained from voxel features through projection operations, where B denotes the batch size, C_0 represents the number of input channels, and H×W indicates the spatial resolution of the BEV feature map. BaseBEV performs multi-stage 2D convolutional encoding of this feature on the BEV plane. By introducing a phased feature extraction mechanism and multi-scale feature fusion, it enhances the BEV features ability to represent objects and spatial structures at different scales.

Specifically, BaseBEV consists of multiple consecutive convolutional encoding stages. The i-th stage takes X_{i-1} as input, performs spatial resolution adjustment through a strided two-dimensional convolution, and further stacks several 3×3 convolutional layers for feature encoding. Its output can be expressed as:

$$X_i = f_i(X_{i-1}), \quad i = 1,2,\dots,L, \tag{1}$$

Where, $f_i(\cdot)$ denotes the overall mapping process of the i-th BEV convolutional module. For its internal single-layer convolutional operation, it can be abstracted as:

$$Z = \sigma\left(BN\left(Conv_{3\times3}(U)\right)\right), \tag{2}$$

Where, U and Z denote input and output features, respectively. $BN(\cdot)$ represents batch normalization, while $\sigma(\cdot)$ denotes the ReLU activation function. Through multi-layer stacking, BaseBEV progressively expands its receptive field within the BEV space, thereby capturing richer contextual information.

Unlike the original SECOND + SECONDFPN, BaseBEV does not rely solely on single-scale features after feature extraction at each stage. Instead, it performs upsampling and scale alignment on BEV features from different stages, followed by fusion in the channel dimension. Let the output feature from the th stage be:

$$U_i = Up_i(X_i) \tag{3}$$

Where, $Up_i(\cdot)$ denotes the corresponding upsampling operation (implemented via transposed convolution or convolution). Subsequently, the features aligned across all scales are concatenated to yield the fused feature representation:

$$U = Concat(U_1, U_2, \dots, U_L) \tag{4}$$

Under certain configurations, BaseBEV further scales and channels the concatenated features to ensure the output matches the input interface of subsequent detector heads, ultimately yielding an enhanced BEV representation:

$$Y = B(X_0) \in \mathbb{R}^{B\times C_{out}\times H\times W} \tag{5}$$

Where, $B(\cdot)$ denotes the overall mapping process of the BaseBEV backbone network, and C_{out} represents the number of channels after multi-scale feature fusion.

Through the above design, BaseBEV implements a modeling process of "phased feature extraction with multi-scale alignment-based feature fusion" in the BEV space. Compared to the original SECOND + SECONDFPN architecture, it achieves more comprehensive integration of spatial and semantic information across different scales. This enhancement endows BEV features with stronger discriminative capabilities in complex road scenes, providing more robust feature inputs for subsequent RPN candidate box generation. Consequently, it effectively boosts overall 3D object detection performance.

3.4. Size-Aware Dual-Branch Attention Mechanism

In sparse point cloud 3D object detection tasks, the strong non-uniformity of LiDAR point cloud distributions leads to significant differences in point cloud density, geometric structure, and observability across objects of varying scales. Particularly in scenarios involving small-scale targets such as pedestrians and cyclists, where target points are sparse and severely affected by occlusion, traditional uniform feature enhancement methods often struggle to simultaneously address the representation needs of both large and small objects. Based on this analysis, this paper introduces a scale-aware dual-branch attention mechanism (SAD-Attention) during the keypoint feature fusion stage of PV-RCNN, as illustrated in Figure 3. This module dynamically distinguishes target scales by estimating voxel density around keypoints. It employs distinct attention branches—a channel attention branch and a spatial attention branch—to differentially enhance keypoint features, thereby improving the models detection capability for multi-scale targets, especially sparse small objects. Let the keypoint features processed by VSA and the fusion layer be denoted as $F \in \mathbb{R}^{B\times C\times N}$, where B represents the batch size, N denotes the number of keypoints, and C indicates the feature dimension of keypoints. The core of SAD-Attention lies in selecting appropriate feature enhancement methods for targets at different scales. Therefore, this paper first estimates the density of keypoints based on the distribution of surrounding voxels, using this as the basis for scale discrimination. Given the set of keypoints $\{P_{b,n}\}_{n=1}^{N}$ in the b(th) sample and the corresponding set of voxel centers$\{v_{b,m}\}_{m=1}^{M_b}$), where $P_{b,n} \in \mathbb{R}^3$, $v_{b,m} \in \mathbb{R}^3$, the number of voxel neighbors surrounding keypoint$P_{b,n}$ can be expressed as:

$$d_{b,n} = \sum_{m=1}^{M_b} \mathbb{I}\left(\left\|p_{b,n} - v_{b,m}\right\|_2 < r\right) \tag{6}$$

Where, r denotes the preset neighborhood radius, and $\mathbb{I}(\cdot)$ represents the indicator function. This paper treats $d_{b,n}$ as an approximate measure of voxel density around keypoints. By applying the threshold r keypoints are classified into "small

target regions" and "large target regions," yielding a size mask:

$$S_{b,n} = I(d_{b,n} < r), S \in \{0,1\}^{B\times N} \quad (7)$$

Where, $S_{b,n} = 1$ indicates that the keypoint belongs to a small object or sparse region, prioritizing channel attention enhancement; $S_{b,n} = 0$ indicates that the keypoint is located in a relatively dense region or large object, prioritizing spatial attention enhancement.

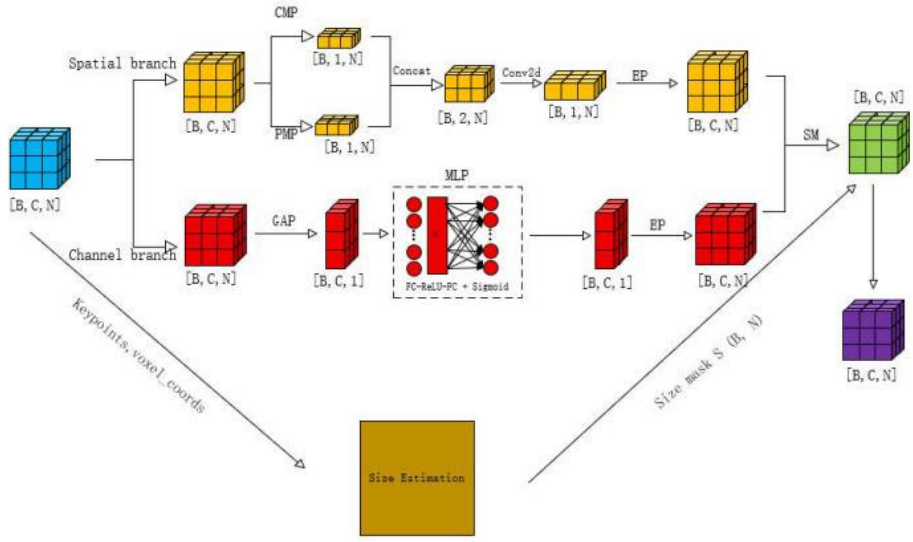

Figure 3. Schematic diagram of the size-aware dual-branch attention mechanism.

During the feature enhancement stage, SAD-Attention constructs two complementary attention branches, as shown in Figure 3: the channel attention branch and the spatial attention branch. The channel attention branch is used to highlight the channel discriminative power of small object features, implemented using a Self-Attention (SE) structure. Specifically, the channel description vector is obtained through global average pooling:

$$z = AP\ (F) \in \mathbb{R}^{B\times C\times 1} \quad (8)$$

and generate channel attention weights through two fully connected layers:

$$a = \sigma(W_2\delta(W_1 z)) \in \mathbb{R}^{B\times C\times 1} \quad (9)$$

Where, $\delta(\cdot)$ denotes the ReLU activation function, and $\sigma(\cdot)$ denotes the Sigmoid function. The final feature representation after channel attention enhancement is:

$$F_{ch} = F \odot a \quad (10)$$

Where, $\odot$ denotes element-wise multiplication with dimension broadcasting.

The spatial attention branch enhances the spatial structural information of large objects. Its core idea is to generate spatial attention maps in the keypoint dimension. First, maximum pooling and average pooling are performed in the channel dimension, yielding:

$$m_{max} = MaxPool_c(F),\ m_{avg} = AvgPool_c(F) \quad (11)$$

Where, $m_{max}, m_{avg} \in \mathbb{R}^{B\times 1\times N}$. Concatenating these two and performing convolution generates the spatial attention weights:

$$b = \sigma(Conv_{k\times k}(Concat[m_{max}, m_{avg}])), b \in \mathbb{R}^{B\times 1\times N} \quad (12)$$

The feature representation after spatial attention enhancement is:

$$F_{sp} = F \odot b \quad (13)$$

After obtaining two enhanced feature streams, SAD-Attention dynamically routes keypoints based on the size mask S. This process can be uniformly represented as:

$$F_{out} = S \odot F_{ch} + (1 - S) \odot F_{sp} \quad (14)$$

Where, S is broadcast across the channel dimension, ensuring each keypoint selects the appropriate attention branch according to its corresponding mask. This design enables small targets (sparsely distributed keypoints) to enhance semantic discrimination through channel attention, while large targets (densely distributed keypoints) strengthen geometric structure representation via spatial attention. This achieves scale-adaptive feature enhancement.

In implementation, this paper reshapes the keypoint fusion features into (B,C,N) after generation, then restores them to (B,N,C) format after SAD-Attention processing to align with the input requirements of the subsequent RoI-grid pooling module. By introducing SAD-Attention, the model achieves multi-scale, differentiated feature modeling at the keypoint level. This significantly enhances feature representation for small objects in sparse point cloud conditions, thereby improving overall detection performance. Ablation experiments further demonstrate that this module complements the BaseBEV backbone effectively, delivering substantial accuracy gains while maintaining network stability.

4. Experiments and Results

4.1. Dataset and Evaluation Metrics

This paper experimentally validates the proposed method on the KITTI 3D Object Detection dataset. The KITTI dataset is one of the most widely used public datasets in the field of 3D object detection, extensively employed for evaluating environmental perception algorithms in autonomous driving scenarios. In the KITTI 3D object detection task, detected objects primarily include three categories: Car, Pedestrian, and Cyclist. Each object in the dataset is annotated as a 3D bounding box to evaluate an algorithms 3D perception capabilities in complex road scenarios.

The KITTI dataset comprises 7,481 annotated training frames and 7,518 test frames. During training, this paper further divides the training samples into a training set (3,712 frames) and a validation set (3,769 frames) according to common classification methods, used for model training and performance evaluation. Samples in the dataset are

categorized into three difficulty levels—Easy, Moderate, and Hard—based on object size, occlusion, and clipping severity, enabling comprehensive evaluation of detection performance across varying complexity levels.

For evaluation metrics, this paper employs Average Precision (AP) as the performance benchmark for 3D object detection, strictly adhering to the official KITTI evaluation protocol. For the 3D detection task, the Intersection over Union (IoU) threshold is set to 0.7 for the Car category and 0.5 for the Pedestrian and Cyclist categories. Detection results are reported separately for each difficulty level to comprehensively evaluate the proposed methods performance across diverse complex scenarios.

4.2. Experimental Setup

This experiment was conducted using the MMDetection3D framework, with all training and testing processes performed on a GPU. The experimental hardware configuration comprised an NVIDIA GeForce RTX 4090 (24GB VRAM), an Intel Xeon E5-2697A v4 CPU, and 64 GB of system memory. The software environment consists of Ubuntu 20.04 operating system, Python 3.8, CUDA 11.3, and PyTorch 1.10.1.

Regarding training strategy, this paper employs an end-to-end optimization approach for the model. The batch size was set to 2, with training conducted over 40 epochs. The AdamW optimizer was used, featuring an initial learning rate of 0.0018 and weight decay of 0.01.

4.3. Experimental Results and Analysis

To validate the performance advantages of the proposed method, comparative experiments were conducted on the KITTI test dataset against several representative 3D object detection methods. Evaluations were performed separately for the BEV detection task and the 3D object detection task, with results shown in Tables 1 and 2, respectively.

Table 1 summarizes the BEV detection performance of each method on the KITTI test dataset. As shown, our method achieves a mAP of 77.33% in the BEV detection task, outperforming all comparison methods. Compared to the baseline PV-RCNN, our method improves the BEV detection mAP by approximately 2.0% and demonstrates stable performance gains across both Moderate and Hard difficulty levels for the Car category.

Furthermore, for the Pedestrian category, our method achieves superior detection results to the comparison methods across all three difficulty levels (Easy, Moderate, and Hard). This demonstrates that the proposed BEV backbone network effectively enhances the representational power of BEV features, while the introduced size-aware attention mechanism helps improve the models discriminative capability for small-sized objects and sparse point cloud scenarios.

Table 1. Results of BEV detection in KITTI test set unit: %

Method	mAP	Car			Pedestrians		
		Easy	Mod	Hard	Easy	Mod	Hard
VoxelNet	60.83	89.35	79.26	77.39	46.13	40.74	38.11
PillarPillars	71.18	90.75	86.92	84.39	60.35	54.45	50.23
PV-RCNN	75.34	94.19	89.99	88.35	66.08	58.89	54.59
EOTL	72.95	91.63	88.19	81.46	66.24	58.11	52.09
EFMFpillars	72.95	91.63	88.19	81.46	66.24	58.11	52.09
PG-RCNN	--	93.39	89.46	86.54	--	--	--
PPDysample	70.49	91.78	87.79	85.08	59.28	51.76	47.28
Our	77.33	93.05	90.38	88.45	70.22	63.43	58.5

Table 2. Results of 3D detection in KITTI test set unit: %

Method	mAP	Car			Pedestrians		
		Easy	Mod	Hard	Easy	Mod	Hard
VoxelNet	51.33	78.47	66.13	57.73	39.48	34.67	31.50
PointPillars	61.87	85.49	75.30	72.37	50.87	45.61	41.52
PV-RCNN	70.98	91.50	82.47	81.98	63.15	55.74	51.05
EOTL	65.56	88.75	76.73	75.31	57.07	50.07	45.44
EFMFpillars	66.03	87.98	77.28	75.57	57.07	51.13	47.20
PG-RCNN	72.21	89.38	82.13	77.33	68.44	60.63	55.36
PPDysample	62.07	87.89	78.13	73.44	50.62	43.28	39.01
Our	74.02	91.94	82.89	82.42	69.42	61.03	56.42

Table 2 presents the 3D object detection results of different methods on the KITTI test dataset. As shown in the table, our method achieves the best performance in the 3D detection task, with an overall mAP of 74.02%. Compared to PV-RCNN, our method improves the 3D detection mAP by approximately 3.0% and demonstrates superior detection accuracy for both vehicle and pedestrian categories.

For the Car category, our method achieves stable improvements across all difficulty levels, with particularly noticeable gains in Moderate and Hard scenarios. This demonstrates the proposed approachs s uperior 3D geometric modeling capabilities in complex road environments. In the Pedestrian category, our method gains significant advantages at Moderate and Hard difficulty levels, validating the effectiveness of the size-aware feature enhancement strategy for handling small-sized and distant objects.

Combining BEV and 3D detection results, our method outperforms existing mainstream approaches across different tasks and difficulty levels. This demonstrates that the BaseBEV backbone network combined with the size-aware dual-branch attention mechanism effectively enhances the expressive power of point cloud features.

4.4. Melting Analysis Experiment

To analyze the specific impact of the two proposed improvement strategies on detection performance, ablation

experiments were conducted on the KITTI test dataset using PV-RCNN as the baseline. The experiments introduced the BaseBEV backbone network and the size-aware dual-branch attention mechanism (SAD-Attention) respectively. The results are shown in Table 3.

When only replacing the backbone network in the baseline method with BaseBEV (Experiment 1), the model shows significant improvements in detection accuracy for the Pedestrian category under Moderate and Hard difficulty levels. This indicates that the improved BEV backbone more effectively models spatial structural information, thereby enhancing detection capabilities for sparse objects.

When only introducing SAD-Attention into the baseline method (Experiment 2), the results demonstrate that this attention mechanism enhances the expressive power of keypoint features, enabling the model to exhibit improved robustness across varying difficulty levels, particularly for small-sized objects and sparse point cloud regions.

When both BaseBEV and SAD-Attention were introduced simultaneously (Ours), the model achieved optimal detection performance across all difficulty levels for both vehicle and pedestrian categories. Compared to experiments with a single module, the combined use of both improvement strategies further enhances overall detection accuracy, with more pronounced performance gains in Moderate and Hard scenarios.

In summary, ablation experiments validate the effectiveness of the BaseBEV backbone and size-aware dual-branch attention mechanism in our method. They demonstrate that integrating these two modules amplifies their respective strengths, providing stable support for overall performance improvement.

Table 3. Results of ablation experiment unit: %

Method	SAD-Attention	BaseBEV	Car-3D			Pedestrians-3D			mAP	
			Easy	Mod	Hard	Easy	Mod	Hard	Car	Pedestrians
PV-RCNN			91.50	82.47	81.98	63.15	55.74	51.05	85.31	56.64
Experiment1		√	91.87	83.54	81.40	66.1	59.54	53.99	85.60	59.97
Experiment2	√		91.82	82.81	82.20	66.85	58.95	54.13	85.61	59.87
Our	√	√	91.94	82.89	82.42	69.42	61.03	56.42	85.75	62.29

5. Conclusion

This paper addresses the limitations of BEV feature representation and the limited adaptability of keypoint features to scale variations in sparse point cloud 3D object detection. We propose SAPV-RCNN, a point-voxel fusion 3D object detection method enhanced with a size-aware attention mechanism and a BEV backbone network. While preserving the overall two-stage detection workflow of PV-RCNN, this method introduces two key improvements: First, replacing the original SECOND and SECONDFPN structures with the BaseBEV backbone to enhance multi-scale feature modeling capabilities in the BEV space; Second, it introduces a scale-aware dual-branch attention mechanism (SAD-Attention) after keypoint feature fusion. This mechanism enables scale-adaptive routing through voxel density estimation and employs both channel attention and spatial attention to enhance keypoint feature representations for objects of varying scales.

Experimental results on the KITTI dataset demonstrate that the proposed method achieves superior performance to the PV-RCNN baseline in both BEV detection and 3D detection tasks, validating the effectiveness and complementary nature of BaseBEV and SAD-Attention. Future work may explore more refined scale estimation strategies and soft routing mechanisms, while also validating the methods generalization capability on larger-scale datasets.

References

[1] Arnold, E., Al-Jarrah, O.Y., Dianati, M., Fallah, S., Oxtoby, D., & Mouzakitis, A. (2019). A Survey on 3D Object Detection Methods for Autonomous Driving Applications. IEEE Transactions on Intelligent Transportation Systems, 20, 3782-3795.

[2] Li, Z., Du, Y., Zhu, M., Zhou, S., & Zhang, L. (2021). A survey of 3D object detection algorithms for intelligent vehicles development. Artificial Life and Robotics, 27, 115 - 122.

[3] Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 3354-3361.

[4] Qi, C., Su, H., Mo, K., & Guibas, L.J. (2016). PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 77-85.

[5] Qi, C., Yi, L., Su, H., & Guibas, L.J. (2017). PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. ArXiv, abs/1706.02413.

[6] Shi, S., Wang, X., & Li, H. (2018). PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 770-779.

[7] Zhou, Y., & Tuzel, O. (2017). VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4490-4499.

[8] Yan, Y., Mao, Y., & Li, B. (2018). SECOND: Sparsely Embedded Convolutional Detection. Sensors (Basel, Switzerland), 18.

[9] Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2018). PointPillars: Fast Encoders for Object Detection From Point Clouds. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12689-12697.

[10] Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., & Li, H. (2019). PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10526-10535.

[11] Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7132-7141.

[12] Woo, S., Park, J., Lee, J., & Kweon, I. (2018). CBAM: Convolutional Block Attention Module. ArXiv, abs/1807.06521.

[13] Yang, B., Luo, W., & Urtasun, R. (2018). PIXOR: Real-time 3D Object Detection from Point Clouds. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7652-7660.

[14] Yin, T., Zhou, X., & Krähenbühl, P. (2020). Center-based 3D Object Detection and Tracking. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11779-11788.

[15] Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., & Han, S. (2022). BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. 2023 IEEE International Conference on Robotics and Automation (ICRA), 2774-2781.

[16] Shi, S., Wang, Z., Wang, X., & Li, H. (2019). Part-A2 Net: 3D Part-Aware and Aggregation Neural Network for Object Detection from Point Cloud. ArXiv, abs/1907.03670.

[17] Shi, S., Jiang, L., Deng, J., Wang, Z., Guo, C., Shi, J., Wang, X., & Li, H. (2021). PV-RCNN++: Point-Voxel Feature Set Abstraction With Local Vector Representation for 3D Object Detection. International Journal of Computer Vision, 131, 531-551.

[18] Yang, Z., Sun, Y., Liu, S., & Jia, J. (2020). 3DSSD: Point-Based 3D Single Stage Object Detector. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11037-11045.