Journal of Electronic Research and Application, 2026, Volume 10, Issue 4
http://ojs.bbwpublisher.com/index.php/JERA
Online ISSN: 2208-3510
Print ISSN: 2208-3502

3D Object Detection Algorithm based on Improved PV-RCNN

Zhicheng Zhao, Shijie Jia*

Dalian Jiaotong University, Dalian 116028, Liaoning, China

***Corresponding author:** Shijie Jia, jsj@djtu.edu.cn

Abstract: In autonomous driving perception, point cloud-based 3D object detection plays an important role. This task still faces two challenges in long-range and small-object detection: loss of fine details and weak context modeling. To solve these problems, this paper proposes HFA-RCNN based on PV-RCNN. The method adds an encoder-decoder structure to the 3D sparse convolution backbone. This design improves multi-scale context modeling and preserves more detailed features. In the BEV feature generation stage, the method also designs a spatial-frequency aggregation network. This network combines complementary information from the spatial domain and the frequency domain. This design improves feature representation. Results on the KITTI dataset show that the proposed method preserves strong detection performance for the Car category and further improves detection accuracy for the Pedestrian and Cyclist categories. These results confirm the effectiveness of the method in long-range and small-object detection.

Keywords: Point cloud; 3D object detection; Autonomous driving

Online publication: May 21, 2026

1. Introduction

The primary objective of this research is to precisely identify and localize objects within a scene by leveraging 3D sensory data. As a core part of autonomous driving perception systems, 3D object detection supplies critical scene information to downstream components, including path planning, decision-making, and risk evaluation, and its effectiveness has a direct impact on the overall safety and robustness of the system. Compared with traditional vision sensors, LiDAR can directly capture the spatial positions and geometric shapes of objects. LiDAR is also less affected by lighting changes, shadows, and complex weather conditions. Consequently, LiDAR-based 3D object detection has become a major research topic in the domain of intelligent perception.

Nevertheless, LiDAR point clouds are characterized by an unordered structure, sparse sampling, and non-uniform spatial distribution, all of which pose substantial challenges to 3D object detection. In autonomous driving scenarios, objects located at greater distances are often represented by only a limited number of points, resulting in weak and insufficient feature representations. Small objects also tend to lose fine details during voxelization, downsampling, and feature aggregation. These problems reduce detection

performance. Therefore, point cloud-based 3D object detection needs to improve long-range context modeling and preserve the fine details of small objects. At the same time, the method also needs to keep good computational efficiency. This goal has become an important research issue in this field.

Existing approaches for object detection in 3D point clouds are mainly classified into two broad groups, namely voxel-driven methods and point-driven methods. Representative voxel-based approaches, including VoxelNet and SECOND, partition point clouds into regular voxel grids and then employ convolutional operations for feature extraction [1,2]. These methods are efficient, but voxelization causes information loss. This problem limits their performance in sparse distant areas and small-object scenes. Point-based methods, such as PointNet and PointRCNN, directly process raw point clouds [3,4]. These methods preserve geometric details better, but they usually require higher computation. They also have limits in feature modeling for large-scale scenes. To balance efficiency and accuracy, PV-RCNN proposes a two-stage detection framework [5]. This framework combines voxel features and point features. It achieves strong performance in 3D object detection tasks.

To combine the efficiency of voxel-based representation with the fine-detail advantage of point-based representation, fusion-based detection frameworks have gradually become an important research direction. PV-RCNN proposes a two-stage 3D object detection method that combines voxel features with keypoint features. The method uses a voxel backbone to extract scene-level semantic information. It then refines proposals through keypoint feature abstraction and RoI-grid pooling. This framework maintains an effective trade-off between detection accuracy and computational efficiency. Nevertheless, PV-RCNN and related approaches still exhibit limitations when dealing with distant and small-scale objects in complex scenes. On the one hand, the receptive field of the traditional sparse convolution backbone is limited. This drawback reduces the network's capacity to model distant contextual dependencies and limits the effective exploitation of scene-level semantics in sparse environments. Meanwhile, the existing feature extraction strategy is still centered primarily on spatial-domain representation learning. The use of multi-scale information and frequency-domain discriminative information is still not enough. This problem restricts the model's ability to detect weak-feature objects, small objects, and boundary details.

In response to these challenges, this study develops a 3D object detection model termed HFA-RCNN, which is built upon PV-RCNN and incorporates a multi-dimensional feature enhancement strategy. The model aims to improve feature representation and detection performance for distant and small objects. First, this paper adds an asymmetric encoder-decoder structure to the 3D sparse convolution backbone. The downsampling process expands the receptive field and gathers richer long-range contextual information. The upsampling process and skip connections then fuse multi-scale features. This design improves context modeling and detail representation in voxel features. Based on this structure, this paper further designs a Spatial-Frequency Aggregation (SFA) module. This module extracts complementary information in parallel from multi-scale convolution in the spatial domain and attention in the frequency domain. It then uses a hybrid attention mechanism for efficient fusion. This design allows the generated Bird's Eye View features to include local details, global structure, and frequency-domain response patterns at the same time. As a result, the model becomes more robust in detecting small and sparse objects. The primary contributions of this study are summarized as follows:

To address the issue of insufficient contextual modeling in PV-RCNN for distant and small-scale object detection, we propose an asymmetric encoder-decoder enhancement structure for the 3D sparse voxel backbone. It improves the feature representation capability of voxel features through multi-scale feature reconstruction.

This paper introduces SFA, which jointly learns object representations from both spatial and spectral perspectives. A hybrid attention mechanism is employed to integrate heterogeneous yet mutually informative features, thereby enhancing the discriminative capability and robustness of BEV representations.

Comparative experiments and visual analysis are conducted using the KITTI benchmark [6]. Experimental findings indicate that the proposed method improves recognition performance for pedestrians and cyclists while maintaining stable results for automobiles, thereby verifying its effectiveness in detecting small and distant targets.

2. Methods

2.1. Overall framework overview

The complete architecture of HFA-RCNN is presented in **Figure 1**. This method uses PV-RCNN as the baseline and follows its two-stage 3D object detection framework. The model mainly consists of four parts: a 3D sparse encoder-decoder backbone, a BEV feature aggregation network, a point-voxel feature fusion module, and an RoI-based proposal refinement head. The input of the network is LiDAR point clouds. After voxelization, the point clouds first enter the 3D sparse convolution backbone for multi-scale feature extraction. The network then generates enhanced Bird's Eye View features for first-stage proposal prediction. In the second stage, the model further refines the proposals through keypoint feature aggregation and RoI-grid pooling. The model then outputs the final object categories and 3D bounding box parameters.

In the first stage, this paper introduces an encoder-decoder structure into the 3D sparse convolution backbone to solve the problems of limited receptive field and weak long-range context modeling in the original backbone. The downsampling process expands the receptive field. The upsampling process and skip connections fuse multi-scale voxel features. This design improves context modeling and detail representation in voxel features. To further improve information use during BEV feature generation, this paper designs SFA. This module models spatial-domain and frequency-domain information in parallel during BEV feature construction. The spatial branch extracts multi-scale local structural features. The frequency branch captures discriminative frequency-domain responses. The fusion module then combines these two types of complementary features efficiently.

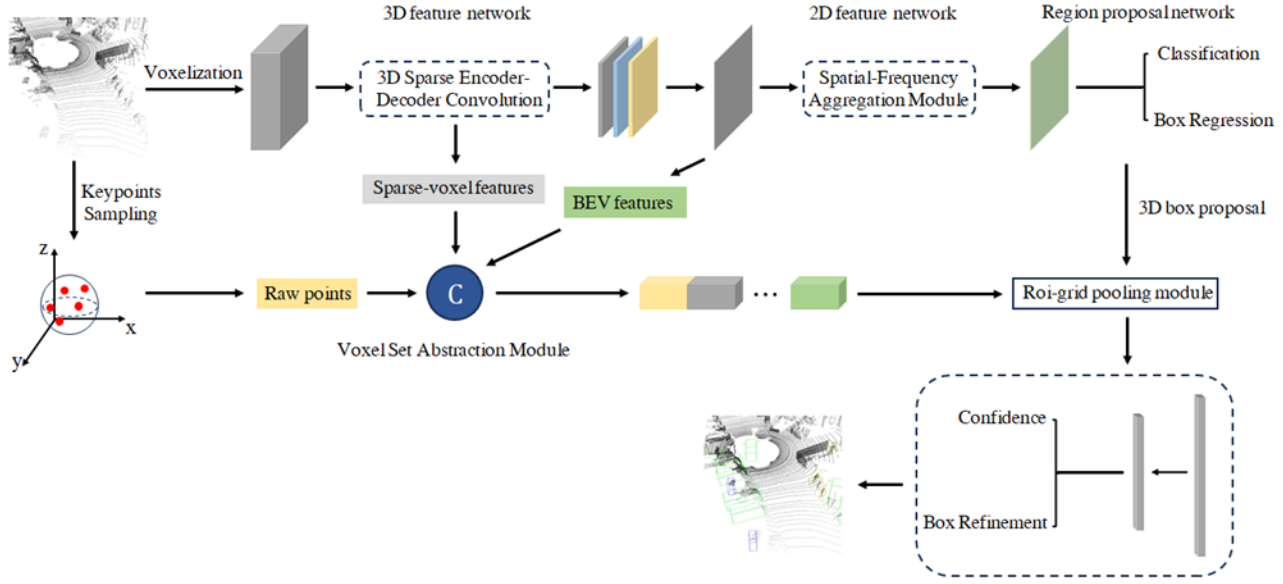


Figure 1. Architecture of HFA-RCNN.

In the second stage, the network follows the point-voxel collaborative feature learning strategy of PV-RCNN. The network first applies farthest point sampling to identify a set of keypoints. Subsequently, the

voxel set abstraction module aggregates contextual information for these keypoints from multi-scale voxel-level features. After that, the network applies RoI-grid pooling to the proposals generated by the region proposal network. This step extracts local features for regular grid points in each proposal region from the keypoint features. Finally, these extracted features are fed into the detection head for object classification and bounding box regression, after which the network produces the final detection outputs.

2.2. Sparse encoder-decoder module

The 3D backbone network in this paper is based on 3D sparse convolution and efficiently extracts hierarchical features from voxelized point clouds. Unlike standard 3D convolution, sparse convolution only performs computation at non-empty voxel locations. It also uses a sparse indexing mechanism to manage valid voxels. Therefore, sparse convolution can greatly reduce computational cost and memory use while keeping feature extraction ability. The resulting representation of a center voxel v is given by:

$$y(v) = \sum_{u \in N(v)} W(u - v) \cdot x(u) + b \tag{1}$$

Here, $N(v)$ represents the local neighborhood of voxel v, $x(u)$ is the feature of the neighboring voxel, W is the convolution kernel weight, and b is the bias term.

The voxel backbone in the baseline PV-RCNN usually consists of stacked sparse convolution blocks. This backbone expands the receptive field and extracts high-level semantic features through progressive downsampling. However, this purely feedforward downsampling structure also causes two problems as follows:

(1) This structure does not make full use of long-range contextual information. As a result, the model finds it hard to capture spatial dependencies in sparse scenes;

(2) Shallow geometric details gradually decrease during continuous downsampling. This problem harms the stable representation of distant and small objects.

To solve this problem, this paper introduces a Sparse Encoder-Decoder module into the 3D backbone network. This module builds a lightweight encoder-decoder structure within a single feature level. It gathers contextual information through downsampling. It restores spatial details through upsampling. It also fuses features from different scales through skip connections. This design improves the multi-scale representation ability of the backbone network without changing the original two-stage detection framework.

As shown in **Figure 2**, given the input voxel feature x_{in}, the Sparse Encoder-Decoder module first refines the same-scale features through two sparse convolutional blocks with stride 1 to obtain an intermediate feature F1.

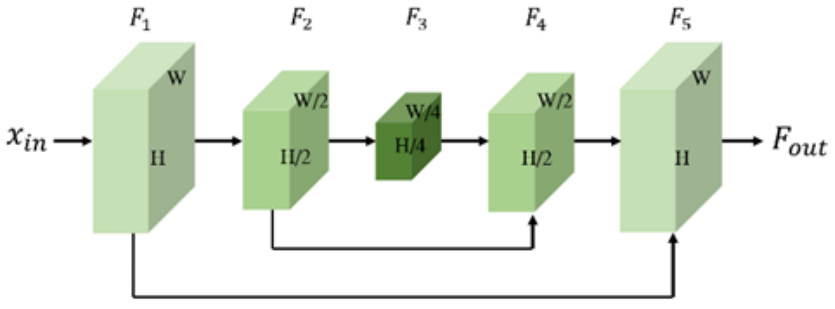


Figure 2. Sparse encoder-decoder structure.

Subsequently, the encoder part performs down sampling using two sparse convolutions with stride 2, Feature F_2 is generated in sequence at half the resolution of the input feature, followed by feature F_3 at one-quarter of that resolution. The above procedure is formulated as:

$$F_2 = Down\left(Block\left(F_1\right)\right) \tag{2}$$

$$F_3 = Down\left(Block\left(F_2\right)\right) \tag{3}$$

Here, *Block*(·) represents the same-scale feature refinement operation using sparse convolution, and Down(·) represents the downsampling operation based on strided convolution. During the encoding process, features F1, F2, and F3 at different scales are preserved for cross-scale fusion in the subsequent decoding stage.

In the decoding stage, the deepest feature F_3 serves as the input. The network gradually restores the spatial resolution through sparse transposed convolution. The network also uses skip connections to fuse this feature with features from the corresponding scales in the encoding stage. In this way, the network obtains the reconstructed features F_4 and F_5. This process is expressed as follows:

$$F_4 = Norm\left(UP\left(F_3\right) + F_2\right) \tag{4}$$

$$F_5 = Norm\left(UP\left(F_4\right) + F_1\right) \tag{5}$$

Here, UP(·) represents the up sampling operation using sparse transposed convolution, and Norm(·) represents the feature normalization operation.

The Sparse Encoder-Decoder module does not replace the original sparse convolution backbone. Instead, this paper adds it as an enhancement unit to key feature layers. This module provides additional modeling for the mid-level and high-level features produced by the backbone. In HFA-RCNN, the Sparse Encoder-Decoder module is introduced after the downsampling operations in the second and third stages of the 3D backbone. The corresponding channel numbers are 32 and 64. This design improves multi-scale context modeling while keeping the extra computational cost under control. The improved 3D backbone keeps the efficiency of sparse convolution. At the same time, it improves both context aggregation and detail recovery. As a result, the backbone provides more discriminative voxel features for later proposal generation and refined detection.

2.3. Spatial-frequency aggregation module

To further improve the representation ability of BEV features, this paper proposes SFA. This module is used in the BEV feature generation stage. This module simultaneously captures spatial-domain and frequency-domain features, thereby strengthening the network's performance in detecting small and distant objects.

As shown in **Figure 3**, given the input BEV feature map X_{bev}, SFA adopts a parallel dual-branch structure for spatial domain and frequency domain feature extraction, respectively. The spatial branch models local structural information and multi-scale spatial relationships to enhance the feature's ability to represent object geometry, boundary details, and surrounding context. The frequency branch captures feature responses across different frequency components to supplement global statistical information and long-range dependency characteristics. Subsequently, a hybrid attention fusion module adaptively aggregates the features from the two branches, generating a more robust BEV representation. This provides effective feature support

for subsequent proposal generation and bounding box regression.

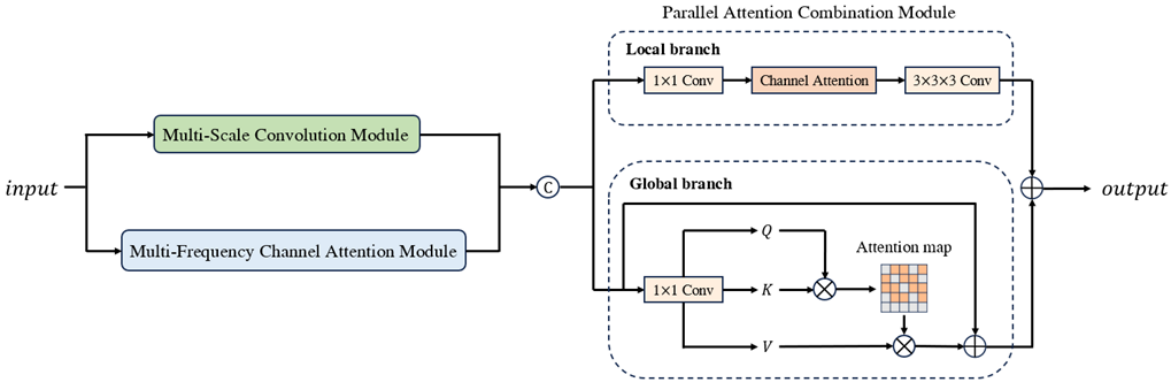


Figure 3. Architecture of the spatial-frequency aggregation network.

2.3.1. Spatial domain feature extraction

To construct multi-scale feature representations for the spatial branch, this study introduces a Multi-Scale Parallel Large Convolution Kernel Module (MSCM). This module adds parallel atrous depthwise convolution branches at different scales. This design improves the ability of BEV features to capture local structures and multi-scale spatial patterns with low computational cost.

The core idea of MSCM is to use atrous convolution to expand the receptive field. Compared with standard convolution, atrous convolution inserts gaps between kernel elements. This design enlarges the effective receptive field without reducing the feature map resolution. Therefore, atrous convolution is more suitable for modeling long-range spatial relationships in dense BEV features. When integrated with depthwise convolution, The module improves multi-level spatial representation learning while keeping the number of parameters and computational cost relatively low.

As shown in **Figure 4**, given the input BEV feature, MSCM first adjusts the channels using a 1×1 pointwise convolution. It then uses a 5×5 standard convolution for initial integration of local neighborhood features to obtain the base feature X_{base}. Building on this, three parallel atrous depthwise convolution branches are constructed, with kernel sizes of 3×3, 5×5, and 7×7, respectively, and a unified dilation rate of d = 3. Through the combination of different kernel sizes and atrous sampling, the three branches acquire spatial receptive fields of different ranges, thereby extracting complementary multi-scale spatial features.

After multi-branch feature extraction, the module concatenates the outputs of all branches along the channel dimension. The module then performs cross-channel nonlinear fusion and dimensionality reduction through a lightweight MLP with a GELU activation function. Finally, the module adds a residual connection and obtains the spatial-domain enhanced feature X_{spatial}. The forward computation process is expressed as follows:

$$X_{\text{base}} = Conv_{5\times5}\left(Conv_{1\times1}\left(BatchNorm\left(X_{\text{bev}}\right)\right)\right) \tag{6}$$

$$X_{\text{multi}} = Concat\left[DWC_{3\times3}^{d=3}\left(X_{\text{base}}\right), DWC_{5\times5}^{d=3}\left(X_{\text{base}}\right), DWC_{7\times7}^{d=3}\left(X_{\text{base}}\right)\right] \tag{7}$$

$$X_{\text{spatial}} = X_{\text{bev}} + MLP\left(X_{\text{multi}}\right) \tag{8}$$

Here, $DWC_{K\times K}^{d}$ represents depthwise atrous convolution with kernel size K and dilation rate d.

Through this coherent and efficient design, the MSCM module endows BEV features with the dual capabilities of local detail expression and global structure perception, without significantly increasing the network parameters.

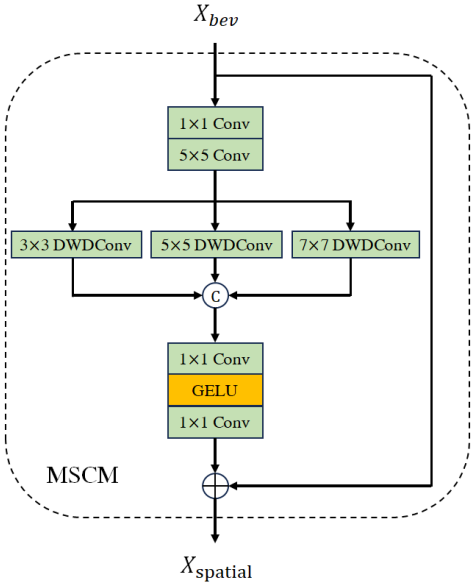


Figure 4. Architecture of the multi-scale convolution network.

2.3.2. Frequency domain feature extraction

To improve the ability of BEV features to model global structure and long-range dependencies, this paper introduces a Multi-Frequency Channel Attention (MFCA) module into the frequency-domain branch. This module maps the input features into multiple frequency subspaces. It then generates channel attention weights from the response information of different frequency components. In this way, the module recalibrates the original features under frequency guidance.

As illustrated in **Figure 5**, let X_{bev} denote the BEV representation provided to the module, adaptive average pooling is first used to compress it into a fixed-size 2D representation $X_{pool} \in R^{C\times H_s\times W_s}$. To balance frequency domain modeling capability with computational cost, this paper sets $H_s = W_s = 7$. On this basis, the feature is decomposed in the frequency domain using 2D discrete cosine transform (2D-DCT). For the k-th frequency component (u_k, v_k), the corresponding 2D-DCT basis function is expressed as:

$$D_{h,w}^{u_k,v_k} = \cos\left(\frac{\pi h}{H_s}\left(u_k + \frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W_s}\left(v_k + \frac{1}{2}\right)\right) \tag{9}$$

Here, (h, w) represent the spatial coordinates, and (u_k, v_k) represent the frequency indices.

To control model complexity while preserving the main frequency information, this paper adopts a Top-K frequency selection strategy, selecting K = 16 representative frequency components from a predefined frequency set for modeling.

For each frequency component, X_{pool} is first multiplied element-wise with the corresponding DCT basis function. The global average, maximum, and minimum values are then computed to obtain the statistical response in that frequency subspace. Subsequently, the statistical results from all selected frequency

components are aggregated to form a multi-frequency channel descriptor vector S. Compared with using a single pooling method, this strategy captures the activation characteristics of channels in the frequency domain from different statistical perspectives, thereby obtaining a more comprehensive frequency domain representation.

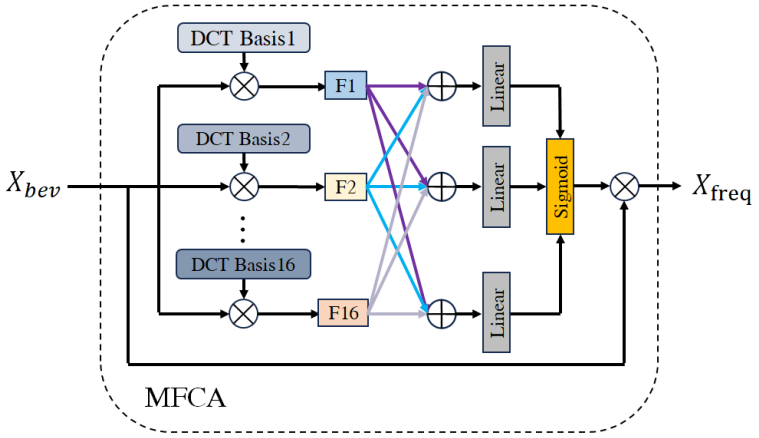


Figure 5. Architecture of the multi-frequency attention network.

The fused frequency domain descriptor vector S is then passed through a lightweight mapping network consisting of two 1×1 convolutional layers. After Sigmoid activation, it generates frequency domain channel weights W_{freq}, which are used to recalibrate the original input features. This computation process can be expressed as:

$$W_{freq} = \sigma\left(MLP\left(S\right)\right) \tag{10}$$

$$X_{freq} = X_{bev} \cdot W_{freq} \tag{11}$$

Here, $\sigma(\cdot)$ represents the Sigmoid nonlinearity, whereas $\cdot$ refers to channel-wise element-wise multiplication.

Through the above design, MFCA explicitly uses response information from different frequency components. This module strengthens channels that are more useful for object detection and suppresses redundant or irrelevant features. This design improves the global representation ability and discriminability of BEV features. It also provides effective frequency-domain feature support for later spatial-frequency fusion.

2.3.3. Parallel attention combination

After the spatial-domain feature $X_{spatial}$ and the frequency-domain feature X_{freq} are obtained, this paper designs a Parallel Attention Combination (PAC) module to further achieve collaborative modeling of cross-domain features. First, the two feature types are concatenated along the channel dimension, after which a 1×1 convolution is applied for channel alignment and initial feature fusion, obtaining a unified feature representation:

$$X_{fuse} = Conv_{1\times1}\left(Concat\left[X_{spatial}, X_{freq}\right]\right) \tag{12}$$

Although X_{fuse} contains both spatial and frequency domain information, relying solely on convolution makes it difficult to simultaneously model global spatial dependencies and select local discriminative features. To address this, PAC uses parallel global and local branches to enhance the fused features.

The global branch models X_{fuse} based on a self-attention mechanism. Specifically, the input feature is first linearly mapped to generate query, key, and value representations. The spatial attention map is obtained by computing the correlation between queries and keys, and the value features are then weighted and aggregated, capturing long-range spatial dependencies. After convolutional mapping and a residual connection, the globally enhanced feature X_{global} is obtained.

The local branch focuses on modeling importance along the channel dimension. First, global average pooling and global max pooling are applied to X_{fuse} to obtain channel descriptors F_{avg} and F_{max}. These are then passed through a shared multilayer perceptron and activated by Sigmoid to generate channel weights:

$$W_{local} = \sigma\left(MLP\left(F_{avg}\right) + MLP\left(F_{max}\right)\right) \tag{13}$$

These weights are subsequently applied to reweight the original features in the channel dimension, yielding a locally refined representation:

$$X_{local} = W_{local} \cdot X_{fuse} \tag{14}$$

Finally, the outputs of the global branch and the local branch are fused element-wise to obtain the output feature of the PAC module:

$$X_{PAC} = X_{global} + X_{local} \tag{15}$$

Through the above design, the PAC module can simultaneously model long-range spatial dependencies and local channel responses based on unified spatial-frequency features, thereby enhancing the global structural representation capability and local discriminative capability of BEV features.

3. Results and discussion

3.1. Experimental dataset

The experimental evaluation in this work is conducted on the KITTI benchmark for 3D object detection, where data were collected in real-world urban scenes with a Velodyne HDL-64E laser scanner. As a standard benchmark widely used in autonomous driving studies, KITTI provides 7,481 samples for training and 7,518 samples for testing. In accordance with the commonly adopted split strategy in KITTI, the official training portion is further divided into 3,712 samples for model training and 3,769 samples for validation. The first subset is used for parameter optimization, while the second is employed for performance evaluation.

The experimental evaluation follows the Average Precision (AP) criterion defined by the KITTI benchmark. Based on the severity of occlusion and truncation, detection tasks are grouped into three difficulty categories, namely Easy, Moderate, and Hard. During evaluation, the 3D IoU threshold is fixed at 0.7 for cars and 0.5 for pedestrians and cyclists. This study reports 3D AP under all three difficulty categories, with the Moderate setting serving as the main basis for method comparison. This evaluation protocol offers a more reliable indication of the overall capability of the model under complicated practical conditions.

3.2. Experimental environment and settings

The method is developed with PyTorch and built upon the OpenPCDet framework. All evaluations are conducted on Ubuntu 20.04 using a computing platform configured with an NVIDIA RTX 3090 graphics card with 24 GB memory. For optimization, the Adam algorithm is employed during model training. The network is trained for 80 epochs with a batch size of 2, and the learning rate, momentum coefficient, and weight decay factor are initialized to 0.01, 0.9, and 0.01, respectively. In addition, multiple augmentation strategies are introduced throughout the training process, including global rotation, scaling, random flipping, and GT-Sampling, in order to improve generalization ability and robustness.

3.3. Analysis of experimental results

To assess the performance of HFA-RCNN, comparative evaluations are conducted on the KITTI benchmark for 3D object detection against several representative approaches. The assessment follows the official AP criterion for 3D detection in KITTI. In the evaluation protocol, the 3D IoU threshold is specified as 0.7 for cars and 0.5 for pedestrians and cyclists. The selected comparison methods include SECOND, AVOD, PointPillars, PointRCNN, Part-A2Net, Pointformer, and the baseline model PV-RCNN [7–10]. The quantitative results are reported in **Table 1**.

Overall, HFA-RCNN achieves strong detection performance on the KITTI dataset. The method shows especially clear advantages for the Pedestrian and Cyclist categories. Compared with the baseline model PV-RCNN, the proposed method improves the AP of the Pedestrian category by 7.36, 6.45, and 4.93% under the Easy, Moderate, and Hard levels, respectively. For the Cyclist category, the proposed method improves the AP by 1.19 and 0.40% under the Easy and Moderate levels, respectively. These results show that the proposed backbone enhancement structure and spatial-frequency aggregation mechanism can better preserve detailed information in sparse point clouds. These results also show that the method improves the model's representation ability for small objects and weak-feature targets.

Table 1. Comparison of 3D detection AP results (%) of different algorithms on the KITTI dataset

Method	Car			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
SECOND	83.34	72.55	65.82	51.07	42.56	37.29	70.51	53.85	46.90
AVOD	81.94	71.88	66.28	50.80	42.81	40.88	64.00	52.18	46.61
PointPillars	82.58	74.31	68.99	51.45	41.92	38.89	77.10	58.65	51.92
PointRCNN	86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
Part-A2 Net	87.81	78.49	73.51	53.10	43.35	40.06	79.17	63.52	56.93
Pointformer	87.13	77.06	69.25	50.67	42.43	39.60	75.01	59.80	53.99
PV-RCNN	90.25	81.43	76.82	55.08	47.93	43.85	85.29	69.29	64.38
Ours	89.08	79.01	77.91	62.44	54.38	48.78	86.48	69.69	63.71
Improvement	-1.17	-2.42	+1.09	+7.36	+6.45	+4.93	+1.19	+0.4	-0.67

For the Car category, HFA-RCNN achieves 77.91% AP at the Hard level. This result is 1.09 percentage points higher than that of PV-RCNN. This finding shows that the proposed method has some advantages in complex scenes, such as distant objects, occlusion, and sparse point clouds. However, at the Easy and Moderate levels, the detection accuracy of the Car category decreases by 1.17 and 2.42%, respectively. When

all categories are considered together, the performance gains of the proposed method mainly appear on small objects and difficult samples. In contrast, the improvement for close-range and large objects is relatively limited. This result is consistent with the design goal of the proposed method. The method mainly focuses on feature enhancement for distant and small objects.

Further comparison with other detection methods shows that HFA-RCNN achieves the best results for the Pedestrian category at all three difficulty levels. HFA-RCNN also remains highly competitive in the Cyclist category. The method outperforms the baseline PV-RCNN at the Easy and Moderate levels. Although the method does not surpass PV-RCNN in the Car category at the Easy and Moderate levels, it significantly improves detection performance for small objects such as pedestrians and cyclists. At the same time, the method maintains comparable overall performance for vehicles. Overall, HFA-RCNN achieves a good balance across different categories. These results confirm the effectiveness of the proposed method in long-range and small-object 3D detection tasks.

3.4. Ablation study

To evaluate the contribution of each improved module to detection performance, this paper conducts ablation experiments on the KITTI dataset with PV-RCNN as the baseline. The experiments examine the Sparse Encoder-Decoder module (SED) and the SFA.

As shown in **Table 2**, both SED and SFA modules improve detection performance for the Pedestrian category. When the model uses only the SED module, the AP at the Easy, Moderate, and Hard levels increases to 60.89%, 52.79%, and 48.57%, respectively. These results show a clear improvement over the baseline. This finding shows that better context modeling and detail representation of 3D voxel features help pedestrian detection. When the model uses only the SFA module, the AP at the Moderate and Hard levels increases to 48.61% and 45.36%, respectively. These results suggest that joint spatial-frequency modeling can improve the discriminability of BEV features to some extent.

Table 2. Effect of different improvement methods on detection accuracy for the pedestrian category

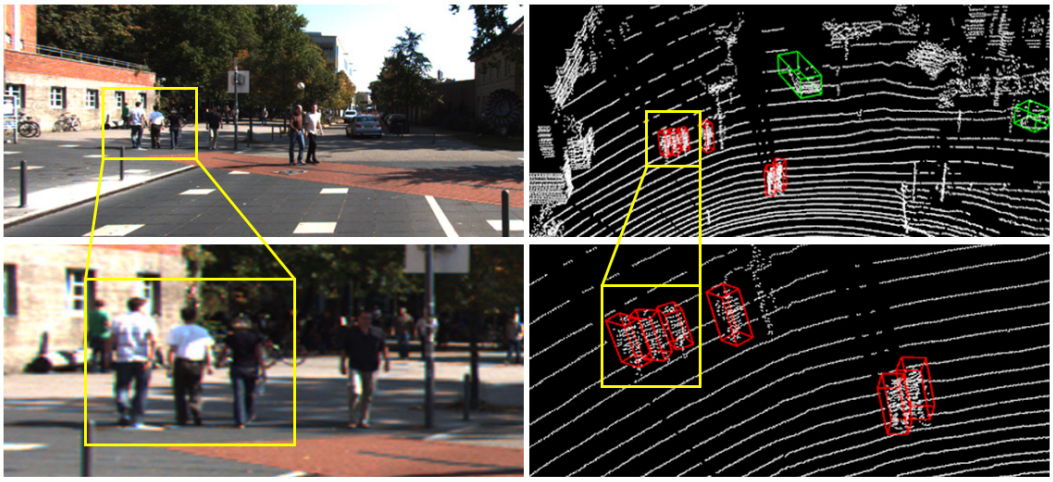
Baseline	SED	SFA	Pedestrian		
			Easy	Moderate	Hard
√			54.93	47.79	44.41
√	√		60.89	52.79	48.57
√		√	54.92	48.61	45.36
√	√	√	62.44	54.38	48.78

When the model uses both the SED and SFA modules, it achieves the best results. AP reaches 62.44%, 54.38%, and 48.78% at the Easy, Moderate, and Hard levels, respectively. These results show that SED and SFA work well together. Their combination further improves the model's detection performance for pedestrian targets.

3.5. Visualization results and analysis

To analyze the detection performance of the proposed HFA-RCNN more clearly, this paper selects typical scenes from the KITTI dataset for visual comparison. **Figure 6** shows the 3D object detection results of the proposed method in complex point cloud scenes. This paper also provides the corresponding images and enlarges the detection details in distant regions. The results show that the proposed method can still detect

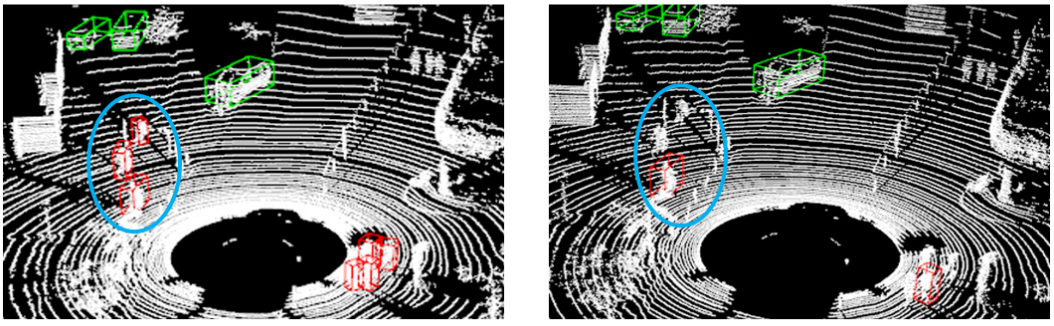
pedestrian targets relatively completely in scenes with complex backgrounds and small objects. This finding demonstrates the good detection ability of the method for sparse point cloud targets. **Figure 7** compares the detection results of the proposed method with those of the baseline model PV-RCNN in the same scenes. The comparison shows that PV-RCNN produces some false detections for pedestrian targets in complex backgrounds. In contrast, the proposed method successfully detects these difficult samples and produces reasonably accurate 3D bounding boxes. These visualization results show that the proposed method achieves more stable detection performance for the Pedestrian category in complex environments. These results also show that the method effectively reduces missed detections.



(a) Point Cloud Detail Display (b) Visualization of the Proposed Algorithm

Figure 6. Visualization of detection results in complex scenes.

(a) Corresponding Image with Point Cloud



(b) Detection Results of PV-RCNN (c) Detection Results of the Proposed Algorithm

Figure 7. Comparison of detection results between the proposed algorithm and PV-RCNN.

4. Conclusion

To improve the representation of distant and small targets in current point-cloud-based 3D object detection, this study proposes HFA-RCNN, a multi-dimensional feature enhancement framework built on PV-RCNN. By introducing a sparse encoder-decoder architecture and a spatial-frequency aggregation network, the method enhances multi-scale contextual modeling and strengthens the discriminability and robustness of BEV features. Experimental results on the KITTI benchmark show improved detection performance for pedestrians and cyclists while maintaining competitive accuracy for cars. Future work will further explore lightweight design and multimodal fusion to enhance efficiency and robustness in practical autonomous driving scenarios.

Disclosure statement

The author declares no conflict of interest.

References

[1] Zhou Y, Tuzel O, 2018, VoxelNet: End-to-End Learning for Point Cloud based 3D Object Detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4490–4499.

[2] Yan Y, Mao Y, Li B, 2018, Second: Sparsely Embedded Convolutional Detection. Sensors, 18(10): 3337.

[3] Qi C, Su H, Mo K, et al., 2017, Pointnet: Deep Learning on Point Sets for 3D Classification and Segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 652–660.

[4] Shi S, Wang X, Li H, 2019, PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 770–779.

[5] Shi S, Guo C, Jiang L, et al., 2020, PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10526–10535.

[6] Geiger A, Lenz P, Urtasun R, 2012, Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3354–3361.

[7] Ku J, Mozifian M, Lee J, et al., 2018, Joint 3D Proposal Generation and Object Detection from View Aggregation, 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 1–8.

[8] Lang A, Vora S, Caesar H, et al., 2019, PointPillars: Fast Encoders for Object Detection from Point Clouds, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12689–12697.

[9] Shi S, Wang Z, Shi J, et al., 2020, Part-A^2 Net: 3D Part-Aware and Aggregation Network for Object Detection from Point Cloud, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13338–13347.

[10] Pan X, Xia Z, Song S, et al., 2021, 3D Object Detection with Pointformer, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7463–7472.