Journal of System Simulation

Volume 35 | Issue 12 Article 16

12-30-2023

Research on 3D Object Detection Method with Cross-module Attention

Renjie Xu
Academy of Army Armored Force, Beijing 100072, China, 1728217581@qq.com

Xiaoming Zhang
Academy of Army Armored Force, Beijing 100072, China

Chen Wang
Zhejiang Sci-Tech University, Hangzhou 310018, China

Peng Wu
Zhejiang Sci-Tech University, Hangzhou 310018, China

Research on 3D Object Detection Method with Cross-module Attention

Abstract

Abstract: To address the issue of feature loss that occurs during the extraction and transmission of target features in 3D object detection tasks using point cloud data, this study proposes an object detection method based on cross-module attention. This method incorporates a channel attention module and a spatial attention module to enhance the crucial feature information. Through feature transformation, the features from different stages of the attention module are connected to mitigate the loss of features during the extraction and transmission process. To tackle the problem of inadequate detection performance in target detection networks for objects of different scales, a cross-scale feature extraction and fusion method is introduced. This method enhances the network's ability to acquire multilevel features by employing multi-scale feature extraction and fusion techniques. Experimental results demonstrate that the proposed method achieves state-of-the-art performance while maintaining a real-time inference speed of 33 Hz.

Keywords

3D object detection, voxel-based network, attention module, multi-scale feature

Recommended Citation

Xu Renjie, Zhang Xiaoming, Wang Chen, et al. Research on 3D Object Detection Method with Crossmodule Attention[J]. Journal of System Simulation, 2023, 35(12): 2680-2691.

第 35 卷第 12 期
2023 年 12 月

系统仿真学报©
Journal of System Simulation

Vol. 35 No. 12
Dec. 2023

Research on 3D Object Detection Method with Cross-module Attention

Xu Renjie[1]，*Zhang Xiaoming*[1]，*Wang Chen*[2]，*Wu Peng*[2]

(1. Academy of Army Armored Force, Beijing 100072, China; 2. Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract： To address the issue of feature loss that occurs during the extraction and transmission of target features in 3D object detection tasks using point cloud data，this study proposes an *object detection method based on cross-module attention*. This method *incorporates a channel attention module and a spatial attention module to enhance the crucial feature information*. Through *feature transformation*，*the features from different stages of the attention module are connected to mitigate the loss of features during the extraction and transmission process.* To tackle the problem of inadequate detection performance in target detection networks for objects of different scales，*a cross-scale feature extraction and fusion method* is introduced. This method *enhances the network's ability to acquire multilevel features by employing multi-scale feature extraction and fusion techniques*. Experimental results demonstrate that the proposed method achieves state-of-the-art performance while maintaining a real-time inference speed of 33 Hz.

Keywords： 3D object detection；voxel-based network；attention module；multi-scale feature

基于跨模块注意力的3D目标检测方法研究

许仁杰[1]，张小明[1]，王晨[2]，吴鹏[2]

(1. 陆军装甲兵学院，北京 100072；2. 浙江理工大学，浙江 杭州 310018)

摘要：针对三维目标检测任务中利用点云数据在提取和传输目标特征过程中发生的特征丢失问题，提出一种*跨模块注意力目标检测方法*。该方法*结合通道注意力模块和空间注意力模块来增强关键特征信息*。通过*特征转换*，*将注意力模块不同阶段的特征连接起来，以减轻提取和传输过程中特征的损失*。针对目标检测网络中不同尺度目标检测性能不足的问题，提出了一种*跨尺度特征提取和融合方法*。*该方法通过采用多尺度特征提取和融合技术增强了网络获取多级特征的能力*。实验结果表明：所提方法在保持33 Hz实时推理速度的同时获得了先进的性能。

关键词：3D目标检测；体素网络；注意力模块；多尺度特征

中图分类号：TP391　　文献标志码：A　　文章编号：1004-731X(2023)12-2680-12

DOI: 10.16182/j.issn1004731x.joss.23-FZ0843E

引用格式：许仁杰, 张小明, 王晨, 等. 基于跨模块注意力的3D目标检测方法研究[J]. 系统仿真学报, 2023, 35(12): 2680-2691.

Reference format: Xu Renjie, Zhang Xiaoming, Wang Chen, et al. Research on 3D Object Detection Method with Cross-module Attention[J]. Journal of System Simulation, 2023, 35(12): 2680-2691.

Received date：2023-07-05　　Revised date：2023-09-11

First author：Xu Renjie(1974-), male, associate professor, doctor, research area: VR and unmanned system simulation.

E-mail: 1728217581@qq.com

0 Introduction

3D object detection is a crucial method for comprehending scenes, whose purpose is to predict the 3D bounding boxes of objects within a scene. It serves as a fundamental element for numerous applications, ranging from autonomous driving and robotics to augmented/virtual reality. In recent approaches, LIDAR is employed to extract precise depths of objects, thereby resolving the issue of limited spatial information provided by images. The position of each point cloud accurately reflects the surface characteristics of actual objects, enabling the acquisition of recognizable 3D representations. This facilitates precise object localization and shape description.

Using point cloud data for learning 3D feature representations can be broadly categorized into two groups: voxel-based methods and point-based methods. Voxel-based methods[1-8] involve converting the raw point cloud into a structured voxel grid representation and extracting the 3D representation using hand-crafted features for each voxel[9-11]. While this approach allows for the extraction of dense and regular 3D features through voxelization and downsampling, it often results in the loss of the spatial 3D structure of the object. In addition, point-based methods[12-18] directly utilize the point cloud to extract features, which enables the acquisition of superior spatial structure features and detection results. However, the processing of large-scale point cloud data is computationally expensive.

This paper presents a 3D object detection framework that integrates an attention mechanism to effectively capture rich feature representations of point cloud data. The framework includes voxel-based and channel-based attention feature encoding networks, as well as a multi-scale feature fusion module[19-20]. To aggregate semantic features of point cloud data, sparse convolutions are employed to generate pseudo-image feature representations. These representations are then combined with attention features to enhance the localization of multi-scale objects within sparse point clouds. Additionally, this paper introduces a detection network that employs a multi-scale feature fusion module. By utilizing the provided pseudo-image feature representation, multi-scale features are extracted and fused to obtain a comprehensive cross-scale feature representation, resulting in improved performance for 3D object detection. Through extensive experiments conducted on the KITTI dataset, the effectiveness and efficiency of the proposed approach, which utilizes attention and multi-scale fusion feature representations, are demonstrated. The main contributions of this paper can be summarized as follows:

The proposed module facilitates the flow of attention across modules, enabling the acquisition of relevant information from both preceding and following modules. As a result, it enables the learning of saliency representation for the object and improves the learning ability of object features. Additionally, the impact of useless features on network detection performance is reduced.

The multi-scale feature fusion module is introduced to perform cross-scale feature fusion on the representation of multi-scale features. This fusion allows for the extraction of more comprehensive multi-level information about the object, which is then utilized for 3D object detection.

To evaluate the effectiveness of the proposed cross-module attention network, experiments are conducted on the KITTI dataset. The results demonstrate that our method surpasses most methods

https://dc-china-simulation.researchcommons.org/journal/vol35/iss12/16
DOI: 10.16182/j.issn1004731x.joss.23-FZ0843E

currently available. Even when compared with state-of-the-art multimodal fusion methods, our method remains highly competitive. Our approach is capable of running in real time, achieving a speed of 33 Hz without compromising detection performance.

1 Proposed Approach

This paper presents a novel single-stage 3D object detection framework called the cross-module attention network. The goal is to improve the accuracy of extracting salient feature representations from point clouds. The proposed network is composed of three main components: a cross-module attention network, a backbone network with multi-scale feature fusion, and an object detection network (Fig. 1). The cross-module attention network performs feature extraction on the collected point cloud and outputs a voxel feature representation in the form of pseudo-images. The sparse point clouds are organized into a voxel grid using the SECOND method[2], resulting in a dense regular 3D voxel feature representation. By connecting the attention modules at different stages, the network enables cross-module feature flow, reducing the loss of the network during feature extraction and transmission and maximizing the potential of the attention modules. Additionally, a multi-scale feature extraction and fusion network is introduced to enhance the performance of the convolutional backbone network by processing the target through multi-scale feature extraction and fusion. The following section provides a detailed description of the cross-module attention network architecture, as well as relevant information about training and inference.

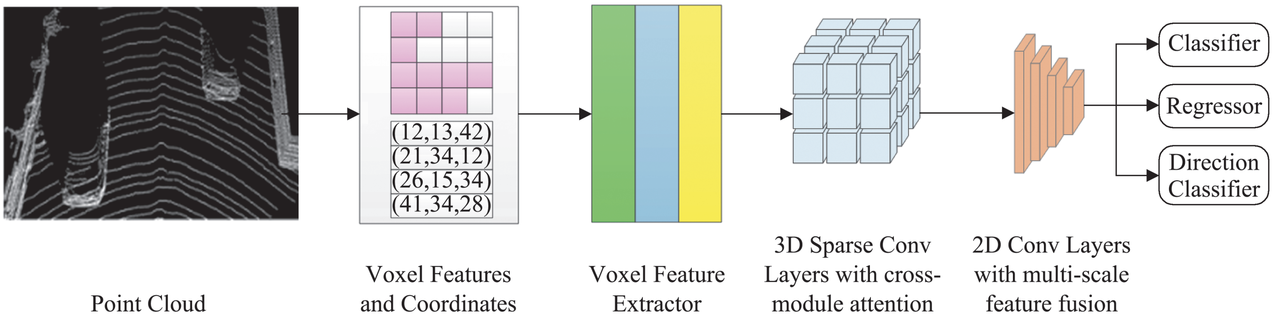

(12,13,42)
(21,34,12)
(26,15,34)
(41,34,28)
Classifier
Regressor
Direction Classifier
Point Cloud
Voxel Features and Coordinates
Voxel Feature Extractor
3D Sparse Conv Layers with cross-module attention
2D Conv Layers with multi-scale feature fusion

Fig. 1 Structure of our proposed detector with cross-module attention

1.1 Cross-module attention network

Point cloud representation

The point clouds obtained from LIDAR form an unordered point set $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_N\}$, of which $\boldsymbol{P}_i \in \mathbb{R}^d$. It includes the point cloud coordinates (x, y, z) represented by $\boldsymbol{d}_i = \mathbb{R}^3$ and other point cloud feature vectors represented by $\boldsymbol{f}_i = \mathbb{R}^{d-3}$. This study establishes the dimensions of the 3D scene where the point cloud is situated within the range of $B \times Q \times D$. The width, height, and length in 3D space are denoted by B, K, and D, respectively.

Voxelization

By setting the size of each voxel to $v_b \times v_q \times v_d$, the size of the voxel grid can be defined to $B' \times Q' \times D'$, where $B' = B/v_b$, $Q' = Q/v_q$, and $D' = D/v_d$. Point clouds are allocated to voxels using the SECOND method, which ensures a uniform distribution. To improve the efficiency of feature extraction, a maximum number of point clouds is set for each voxel. The resulting information includes the voxel coordinates and the point clouds contained within each voxel. For the voxel-based feature encoding, a small PointNet[17] is employed. This encoder takes a collection of voxels and the associated features of all point clouds within each voxel, resulting in $C\times N\times T$ outputs. Here, N represents the total number of voxels and T represents the number of point clouds in each voxel. Subsequently, a max-pooling operation is performed on the point clouds within each voxel, yielding a voxel feature representation of size $C\times N$. The feature of each voxel is represented as a vector of size $C\times 1$. Finally, the obtained features are tiled, and both the voxel features and pointwise features are concatenated to obtain the voxel features.

Cross-module attention module

The cross-module attention network facilitates information flow by connecting different attention modules. After the feature learning process of the attention module in the previous stage, the current features obtained are transferred to the attention module in the next stage. This allows the attention modules in different stages to cooperate and exchange feature information, thereby enhancing the feature learning ability of the attention module and maximizing the role of the attention mechanism.

Fig. 2 illustrates the structure of a cross-module attention network: The white modules represent down-sampling sparse convolutional layers; the blue modules represent size-invariant sparse convolutional layers, and the green module represents sparse convolutional output layers. The network comprises two types of sparse convolutions, with each stage consisting of a normal sparse convolution module and two submanifold sparse convolution modules. Down-sampling is performed on the Z-axis, and the resulting features are then passed through a cross-module attention network to obtain a discriminative feature representation of the object. By converting sparse data into dense feature maps, point cloud data are mapped into pseudo-image feature maps.

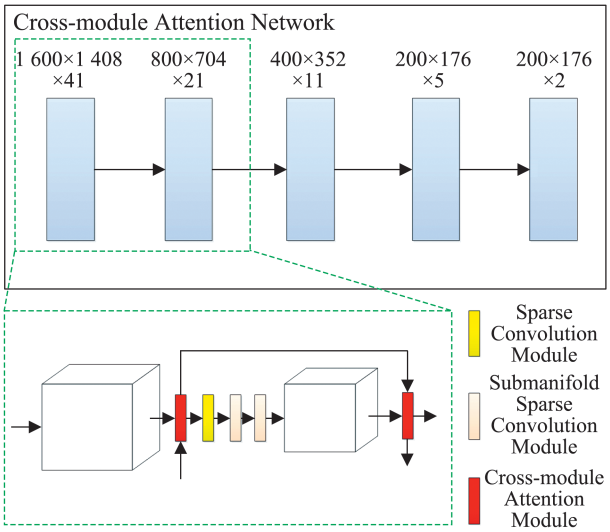

Cross-module Attention Network
1 600×1 408
×41
800×704
×21
400×352
×11
200×176
×5
200×176
×2
Sparse Convolution Module
Submanifold Sparse Convolution Module
Cross-module Attention Module

Fig. 2 Voxel feature extraction network based on cross-module attention

Since the point cloud is sparse and unevenly distributed, the number of points for a voxel may vary. Points within each voxel are collected to form input features. T and C represent the maximum number of points per voxel and the number of channels per point feature, respectively. Voxel grid $\boldsymbol{V}$ consists of k voxels, with the voxel grid defined as $\boldsymbol{V}=\{\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_k\}$, where $\boldsymbol{V}_k\in\mathbb{R}^{N\times C}$ represents the k-th voxel of $\boldsymbol{V}$. The cross-module attention module is shown in Fig. 3. Given a voxel $\boldsymbol{V}_k$, the point cloud feature across channels is aggregated by performing average pooling to obtain its response $\boldsymbol{E}_k\in\mathbb{R}^{1\times C}$. Subsequently, a fully connected layer is utilized to encode the feature response $\boldsymbol{S}_k=\delta(w_1\boldsymbol{E}_k)$, where w_1 denotes the weight parameter of the fully connected layer; δ denotes the ReLU activation function, and $\boldsymbol{S}_k\in\mathbb{R}^{1\times C}$ signifies the channel attention of the current module V_k. The resulting cross-module attention representation, noted as $A_i=\delta(w_2(S_k \oplus A_{i-1}))$, is obtained by connecting it with the channel attention of the previous module. $\oplus$ denotes channel-wise splicing, and $A_i \in \mathbb{R}^{1\times C}$ represents the resulting cross-module channel attention. Channel-based voxel feature $F_1=\sigma(V \otimes A_i)$, and σ are Sigmoid function.

1.2 Multi-scale feature fusion network

This paper utilizes a 2D convolutional backbone network to extract and integrate features more effectively. The backbone network consists of two main components: a network for extracting multi-scale features and a multi-scale feature fusion network for mapping and merging cross-scale features. Due to the significant variations in the scales of 3D objects in real-world environments, cross-scale features assist in addressing the scale differences in 3D object detection. Even though feature pyramid networks (FPNs) are unable to perceive scales well, a method is introduced in this research to obtain scale features through cross-scale fusion. Building upon the FPN framework, the 3D multi-scale feature representation is extracted, and the multi-scale feature representation is fused across scales, as illustrated in Fig. 4.

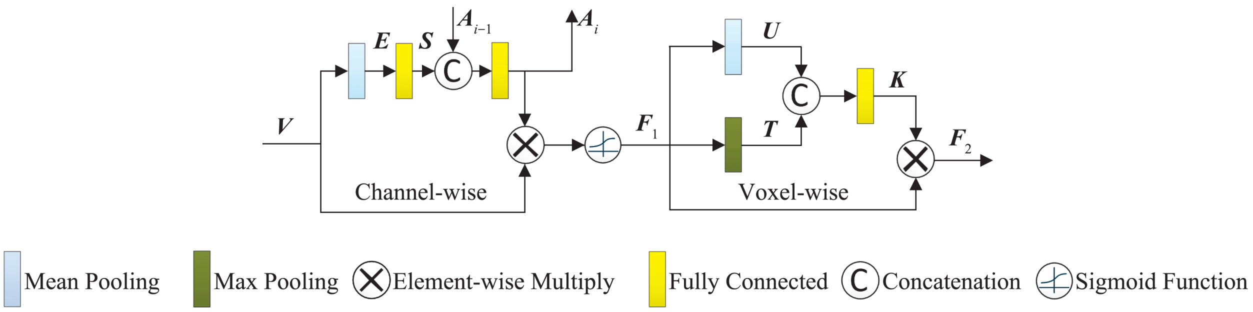


Fig. 3　Cross-module attention module

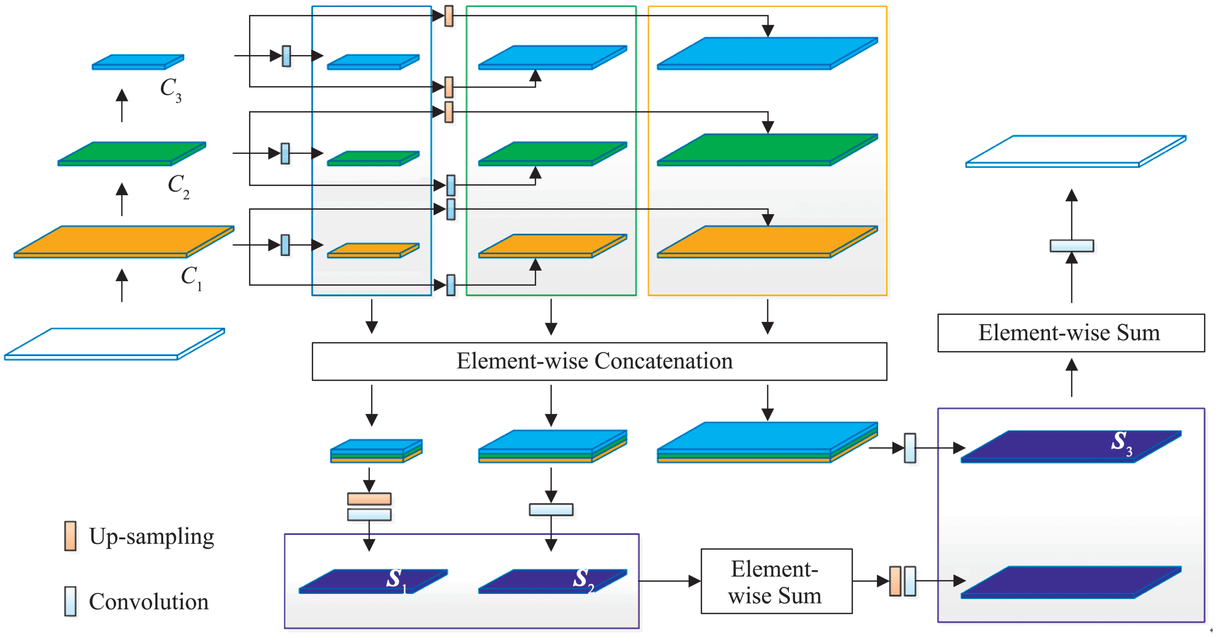


Fig. 4　Multi-scale feature fusion network

Multi-scale feature extraction network

In this work, the pseudo-image features are generated by the voxel feature extraction layer and used as input. The 3D scale feature then undergoes convolution down-sampling. Specifically, as shown in the left part of Fig. 3, the down-sampling operation produces three hidden feature representations denoted as C_1, C_2, and C_3, with corresponding block shapes of (C_0, Q, B), $(2C_0, Q/2, B/2)$, and $(4C_0, Q/4, B/4)$, respectively. By utilizing the pyramid sampling module, cross-layer feature maps are generated, as illustrated in the upper part of Fig. 3. High-level features offer larger receptive fields and richer semantic information, while low-level features maintain higher resolution. Therefore, cross-layer feature mapping effectively captures multi-level information, providing a more comprehensive and salient feature representation for the object. Specifically, the feature pyramid $\{C_1^1, C_1^2, C_1^3\}$ is extracted based on C_1, where C_1^1 has the same size as C_1. C_1^2 and C_1^3 are obtained by performing two down-sampling operations on C_1, respectively. C_1^2 has the same size as C_2, and C_1^3 has the same size as C_3. Similarly, feature pyramid $\{C_2^1, C_2^2, C_2^3\}$ is extracted based on C_2, and C_2 is up-sampled and down-sampled to obtain C_2^1 and C_2^3, respectively. Finally, $\{C_3^1, C_3^2, C_3^3\}$ is obtained by two up-sampling operations based on C_3.

Multi-scale feature fusion network

The resulting features are then inputted into a multi-scale feature fusion network. To fully leverage cross-scale features, as depicted in Fig. 3, C_1^i, C_2^i, and C_3^i are connected along the channel to connect the features of the same scale. A series of convolution operations or up-sampling operations are then performed to up-sample the output of each scale, generating feature maps of the same size. Multiple scales are processed to generate a feature map $S=\{S_1, S_2, S_3\}$, and these feature maps are concatenated into one feature map, resulting in a feature map of size $(4C_0, Q/4, B/4)$. Subsequently, these features are passed through a convolutional layer, and an activation function is applied to generate a feature map of size (C_0, Q, B). Finally, three convolutions of 1×1 are applied to predict the classification, regression, and orientation, respectively. Once more, it is worth noting that high-level features offer larger receptive fields and richer semantic information, whereas low-level features maintain higher resolution. Hence, cross-scale feature mapping efficaciously captures multi-level information, providing a more comprehensive and richer feature representation for the object.

1.3 Object detection network and loss function

For regression tasks, similar to voxel-based methods in other literature, the proposed 3D object detection network uses the same locational loss function proposed in SECOND[2], and the remaining object can be encoded by the following formula:

$$\begin{aligned}&\Delta x=\frac{x_{\mathrm{g}}-x_{\mathrm{a}}}{d_{\mathrm{a}}},\Delta y=\frac{y_{\mathrm{g}}-y_{\mathrm{a}}}{d_{\mathrm{a}}},\Delta z=\frac{z_{\mathrm{g}}-z_{\mathrm{a}}}{d_{\mathrm{a}}}\\&\Delta w=\log_2\frac{w_{\mathrm{g}}}{w_{\mathrm{a}}},\Delta l=\log_2\frac{l_{\mathrm{g}}}{l_{\mathrm{a}}},\Delta h=\log_2\frac{h_{\mathrm{g}}}{h_{\mathrm{a}}}\\&\Delta\theta=\sin(\theta_{\mathrm{g}}-\theta_{\mathrm{a}})\end{aligned}\qquad(1)$$

where $d_{\mathrm{a}}=\sqrt{(l_{\mathrm{a}}^2+w_{\mathrm{a}}^2)}$; x, y, and z are the center coordinates of the box; w, l, and h are the size of the box, corresponding to the width, length, and height, respectively; θ is the heading angle of the box; Subscripts a and g are used to represent the anchor and ground-truth box, respectively.

https://dc-china-simulation.researchcommons.org/journal/vol35/iss12/16
DOI: 10.16182/j.issn1004731x.joss.23-FZ0843E

The loss of the regression task is determined by Eq. (2):

$$L_{reg}=\sum_{res\in(x,y,z,w,l,h)}\text{Smooth}L1(\Delta res) \tag{2}$$

For the branches of detection classification, the imbalance of positive and negative samples is treated with the focus loss proposed by RetinaNet:

$$L_{\text{cls}}=-\alpha_t(1-p_t)^{\gamma}\log_2(p_t) \tag{3}$$

where p_t is the estimated probability of the model; α and γ are the focus loss functions. α=0.25 and γ=2 are adopted during training.

The voxel network directly predicts the radian offset, but in the case of the box of 0 radians and π radians, there will be conflict. Even if the boxes corresponding to these two angles are completely coincident, it will still cause a lot of loss. The problem is solved by introducing a new direction loss:

$$L_{\text{dir}}=\text{Smooth}L1(\Delta\theta) \tag{4}$$

In Eq. (4), by adding a direction regressor using the softmax loss function to the output, the boxes with opposite directions are regarded as the same object, and the conflict between 0 and π directions is solved.

In summary, the total loss is as follows:

$$L_{\text{total}}=\beta_0 L_{\text{cls}}+\frac{1}{N_{\text{pos}}}(\beta_1 L_{reg}+\beta_2 L_{\text{dir}}) \tag{5}$$

where $\beta_0=1.0$, $\beta_1=2.0$, and $\beta_2=0.2$; N_{pos} is the number of positive anchors. A relatively small value of β_2 is used to avoid situations where the network has difficulty in identifying the orientation of objects.

2 Experiment

2.1 Experiment detail

Experimental dataset and evaluation metrics

The proposed network is trained on the KITTI dataset[21] for 3D object detection and bird's eye view object detection including cars, pedestrians, and cyclists. The dataset is divided into training and validation sets following the protocol, with 3 712 samples used for training and 3 769 samples used for validation. Each class is evaluated based on three levels of difficulty (easy, medium, and hard) determined by object height, occlusion, and truncation degree in the 2D results. In accordance with the official KITTI evaluation protocol, the IOU threshold is set to 0.7 for cars and 0.5 for pedestrians and cyclists. This threshold is consistently applied to both 3D object detection and bird's-eye view object detection, while AP is used as a metric for comparing different methods.

Training

Regarding the training process, this paper employs the Adam optimizer, with a batch size of 4 and all models trained for 100 epochs. The initial learning rate is set to 0.003, and a weight decay factor of 0.01 is applied. Data augmentation techniques are utilized during training to address the issue of limited real data leading to poor network performance. Specifically, a database is created from the training dataset, containing the labels of all ground-truths and their associated point cloud data (i.e., points inside the ground-truth 3D bounding box). Global scaling and rotation are applied to the entire point cloud and all ground-truth regions, where a scaling factor is randomly selected in the range of [0.95, 1.05] and a rotation angle is randomly selected from the direction of [−π/4, π/4]. The network is trained on an Intel Core i9-10900X CPU (four cores), 16 GB of DDR4 memory, and an RTX 3090 GPU. It takes 15 h to train a point cloud object detection network with an RTX 3090 GPU.

Experiment detail

For the detection of cars and other related classes of objects, the point cloud is cropped in the range of [−3, 1], [−40, 40], and [0, 70.4] along the *x*, *y*, and *z* axis. The voxel size of $v_d=0.1$, $v_q=0.05$, and $v_b=0.05$ are used to partition the point cloud space into uniformly distributed voxel units for all tasks. Each voxel is randomly sampled with an equal number of points, and the maximum number of points per voxel is set to 5. As the size of the objects to be detected is generally fixed, fixed-size anchors are defined based on the size and center position of all real objects. In the KITTI training dataset, the rotation angle ranges from 0 to 1.57. For cars, an anchor size of B = 1.6, D = 3.9, and Q = 1.56 is used, with a center at $Z = -1.78$. For pedestrians, an anchor size of B = 0.6, D = 0.8, and Q = 1.73 is used, and for cyclists, an anchor size of B = 0.6, D = 1.76, and Q =1.73 is employed, with both centers set to $Z = -0.6$.

2.2 Experiment result

In this paper, the model is compared with a voxel-based method, especially with the pseudo-image representation method. For comparison, this section reports the results of point-based and other voxel-based methods as shown in Table 1 and Table 2.

Table 1 Performance of bird's eye views on KITTI validation set

Method	*t*/s	Car			Pedestrian			Cyclists		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
F-PointNet [12]	0.17	88.70	84.00	75.33	58.09	50.22	47.20	75.38	61.96	54.68
AVOD [22]	0.08	86.80	85.44	77.73	42.51	35.24	33.97	63.66	47.74	46.55
MV3D [9]	0.24	85.82	77.00	68.94	N/A	N/A	N/A	N/A	N/A	N/A
SECOND [2]	0.05	88.07	79.37	77.95	55.10	46.27	44.76	73.67	56.04	48.78
VoxelNet [3]	0.23	89.35	79.26	77.39	46.13	40.74	38.11	66.70	54.76	50.55
ours	0.03	**89.99**	**87.39**	**85.79**	**61.07**	**55.26**	**51.09**	**84.87**	**70.55**	**66.74**

Table 2 Performance of 3D object detection on KITTI validation set

Method	*t*/s	Car			Pedestrian			Cyclists		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
F-PointNet [12]	0.17	81.20	70.39	62.19	51.21	44.89	40.23	71.96	56.77	50.39
AVOD [22]	0.08	73.59	65.78	58.38	38.28	31.51	26.98	60.11	44.90	38.80
MV3D [9]	0.24	66.77	52.73	51.31	N/A	N/A	N/A	N/A	N/A	N/A
SECOND [2]	0.05	83.13	73.66	66.20	51.07	42.56	37.29	70.51	53.85	46.90
VoxelNet [3]	0.23	77.47	65.11	57.73	39.48	33.69	31.51	61.22	48.36	44.37
ours	0.03	**87.19**	**77.45**	**75.67**	**54.94**	**49.61**	**45.56**	**80.34**	**64.83**	**62.05**

Table 1 shows the bird’s eye view object detection results. Our model outperforms other methods in detecting cars and cyclists. The mAP for cars at different difficulty levels is 89.99% (easy), 87.39% (moderate), and 85.79% (hard). All difficulty levels achieve the best performance, with easy samples outperforming the SECOND network by 1.92%, moderate samples leading by 8%, and hard samples leading by nearly 7.8%. For cyclists, the mAP at all difficulty levels is 84.87% (easy), 70.55% (moderate), and 66.74% (hard), all surpassing the VoxelNet network by at least 15%. Similarly, Table 2 presents the 3D object detection results, where our model also demonstrates superior performance in detecting cars and cyclists. The 3D mAP for cars at different difficulty levels is 87.19% (easy), 77.45% (moderate), and 75.67% (hard). Easy, moderate, and hard samples achieve the best performance, with 3D mAP

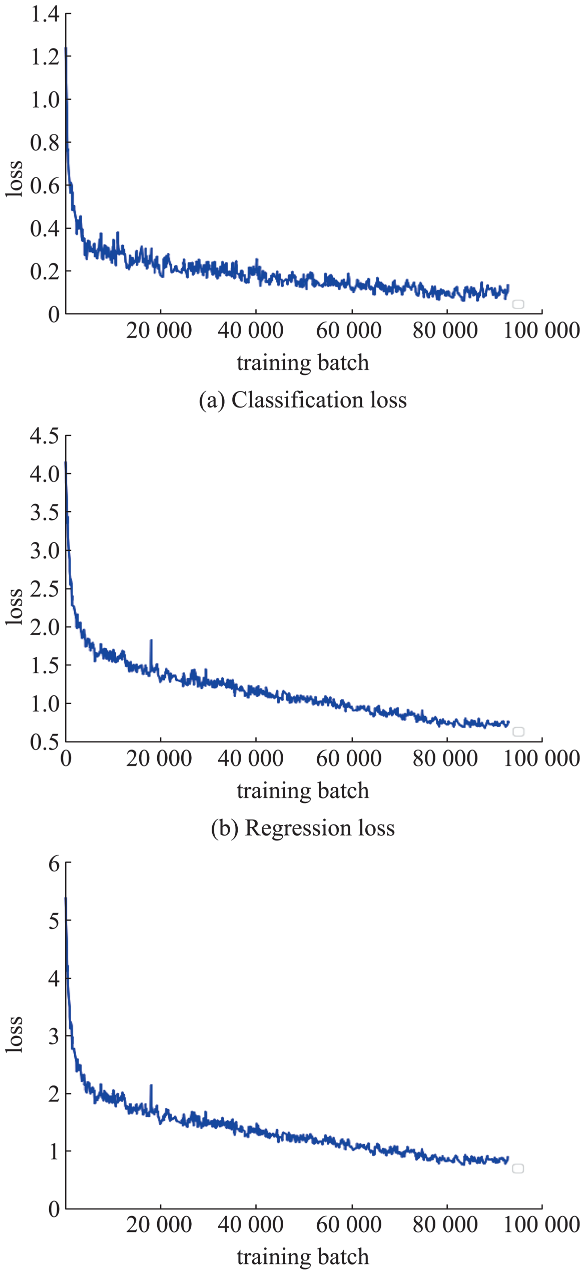
https://dc-china-simulation.researchcommons.org/journal/vol35/iss12/16
DOI: 10.16182/j.issn1004731x.joss.23-FZ0843E

of 80.34%, 64.83%, and 62.05% for cyclists in all difficulty levels, all leading the VoxelNet network by at least 13%. Our model exhibits faster inference speed compared to other voxel-based methods while providing competitive results. Moreover, it demonstrates excellent detection performance across all difficulty levels, indicating a well-balanced scheme in terms of speed and accuracy.

Among the methods with published results, our proposed method achieves the best performance in both BEV and 3D tasks. The evaluation results on the validation set demonstrate that our method can produce highly accurate detections with fast inference speed. The cross-module attention mechanism generates high-quality 3D detection proposals for all object classes and ensures accurate detection even in scenes with dense point clouds or objects.

To evaluate the proposed model, Fig. 5 shows the losses and the total losses for each branch. We plot the training loss using the loss at the training time and evaluate the indicator changes with the training batches. It can be observed that the training loss gradually decreases with an increase in the number of batches fed and shows an upward and downward trend in the process.

For better visualization, Fig. 6 displays the 3D bounding box and predicted confidence score generated by the cross-module attention network. The bounding boxes are shown in green for car predictions, yellow for cyclists, and blue for pedestrians. The visualization demonstrates that our model provides highly accurate 3D bounding boxes for all object classes and effectively detects salient objects, reducing the interference of noise in the network.

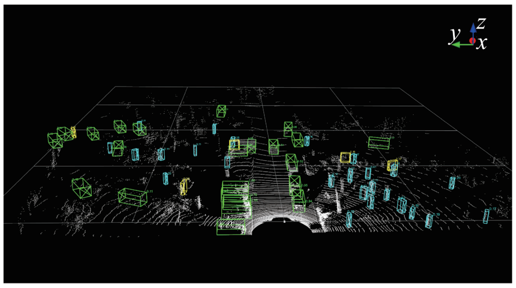


(a) Classification loss

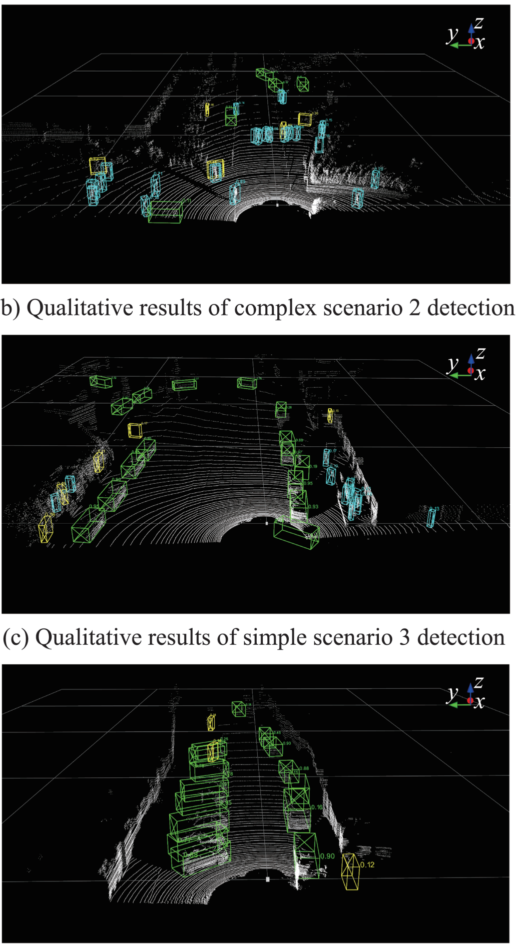
(b) Regression loss

(c) Total loss

Fig. 5 Training loss for each branch



(a) Qualitative results of complex scenario 1 detection



(b) Qualitative results of complex scenario 2 detection

(c) Qualitative results of simple scenario 3 detection

(d) Qualitative results of simple scenario 4 detection

Fig. 6 Qualitative results for car, cyclist, and pedestrian detection on KITTI validation set

2.3 Ablation study

An ablation study is conducted on the KITTI validation set to analyze the impact of different components on the overall performance. The study involves the evaluation of various modules and their effects on the accuracy of the car class and cyclists. A1 represents attention encoding based on voxels and channels in the first sparse convolution block; A2, A3, and A4 represent attention encoding in the second, third, and fourth sparse convolution blocks, respectively; multi-scale features (MSF) refer to the execution of multi-scale feature fusion operations on pseudo-image features; cross-module attention (CMA) refers to the establishment of cross-module connections on attention modules.

The ablation analysis of different parts of the model is shown in Table 3, using the average BEV accuracy of the car class as the evaluation criterion. In the first row of Table 3, the baseline control sample is the basic SECOND network, which serves as a benchmark for ablation experiments. We remove the CMA and MSF modules from our model as the baseline and achieve a 3D mAP of 89.96%, 87.07%, and 79.66%, respectively. To improve the performance in detecting difficult samples and occluded objects, voxel- and channel-based attention encoding is applied to A2 and A3, resulting in moderate and hard sample performance of 87.1% and 84.45%, respectively. Further incorporating voxel- and channel-based attention encoding into A2, A3, and A4 improves the performance of hard samples to 84.9%. As can be seen from the first three rows, by performing voxel-and channel-based attention encoding on sparse convolution blocks, the network’s detection of difficult samples can be significantly improved, and the detection performance of occluded objects can be improved. This shows that voxel- and channel-based attention encoding can obtain more distinctive 3D feature representations, especially for small-scale objects or occluded objects with sparse points; the model in the fourth row adds multi-scale feature fusion to the baseline model in the first row. The performance is improved to 87.23% and 82.54% for moderate and hard samples, respectively. The fifth row performs more cross-module attention connection operations than the third row. It can be found that the accuracy of easy and medium samples is improved by 0.36 and 0.28, respectively, but the accuracy of difficult samples has decreased. The sixth-to-eighth rows add multi-scale feature fusion modules, and the effect of the MSF module is further explored. It can be noted that the detection performance of the network has been significantly improved with the addition of the MSF module. Compared with the network without the MSF module, the optimal detection performance has been improved by 0.48%, 0.4%, and 1.52% for the easy, moderate, and hard samples, respectively. The MSF module can provide robust and discriminate feature representations and make full use of them to estimate 3D bounding boxes.

Table 3 Ablation experiments on KITTI validation set for car in bird's eye views %

A1	A2	A3	A4	MSF	CMA	Easy	Moderate	Hard
						89.96	87.07	79.66
	√	√				89.82	87.10	84.45
	√	√	√			89.15	86.71	84.90
				√		89.75	87.23	82.54
	√	√	√		√	89.51	86.99	84.27
		√	√	√	√	89.70	87.17	85.60
	√	√	√	√	√	**89.99**	**87.39**	**85.79**
√	√	√	√	√	√	89.73	87.25	85.49

The ablation analysis of different parts of the model is shown in Table 4, using the average BEV accuracy of the cyclists as the evaluation criterion. Incorporating voxel- and channel-based attention encoding into A2 and A3 improves the performance of moderate and hard samples to 68.32% and 64.53%, respectively. With voxel-and-channel attention encoding in A2, A3, and A4, all samples' performance is improved to 82.82%, 70.36%, and 66.22%, respectively. The model in the fourth row adds multi-scale feature fusion to the baseline model in the first row, and all samples ' performance is improved to 83.12%, 68.51%, and 63.34%, respectively. The sixth-to-eighth rows add multi-scale feature fusion modules, and the optimal detection performance has been improved by 4.1%, 3.48%, and 4.75% for the easy, moderate, and hard samples, respectively, compared with the network without the MSF module. These results clearly illustrate the effectiveness and robustness of integrating attention encoding, CMA modules, and MSF modules in the model.

Table 4 Ablation experiments on KITTI validation set for cyclists in bird's eye views %

A1	A2	A3	A4	MSF	CMA	Easy	Moderate	Hard
						82.51	66.15	61.28
	√	√				79.25	68.32	64.53
	√	√	√			82.82	70.36	66.22
	√	√	√		√	80.77	67.07	62.00
				√		83.12	68.51	63.34
		√	√	√	√	83.44	68.07	63.58
	√	√	√	√	√	**84.87**	**70.55**	**66.75**
√	√	√	√	√	√	82.16	68.62	64.04

3 Conclusions

In this study, a novel cross-module attention network is proposed for 3D object detection in point clouds. Leveraging the attention mechanism, the cross-module attention network aggregates the voxel grid into unified features and projects them into pseudo-image feature representations, enhancing the key information of the object while suppressing interference. Furthermore, the backbone network incorporates multi-scale feature fusion, which fuses cross-scale features of pseudo-images to generate compact representations of different classes, enabling more accurate detection. Experimental results demonstrate that the proposed method achieves state-of-the-art performance at real-time speed.

References：

[1] Lang A H, Vora S, Caesar H, et al. PointPillars: Fast Encoders for Object Detection from Point Clouds[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 12689-12697.

[2] Yan Yan, Mao Yuxing, Li Bo. SECOND: Sparsely Embedded Convolutional Detection[J]. Sensors, 2018, 18 (10): 3337.

[3] Zhou Yin, Tuzel O. VoxelNet: End-to-end Learning for Point Cloud Based 3D Object Detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 4490-4499.

[4] He Chenhang, Zeng Hui, Huang Jianqiang, et al. Structure Aware Single-stage 3D Object Detection from Point Cloud[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 11870-11879.

[5] Simon M, Milz S, Amende K, et al. Complex-YOLO: An Euler-region-proposal for Real-time 3D Object Detection on Point Clouds[C]//Computer Vision-ECCV 2018 Workshops. Cham: Springer International Publishing, 2019: 197-209.

[6] Yang Bin, Luo Wenjie, Urtasun R. PIXOR: Real-time 3D Object Detection from Point Clouds[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 7652-7660.

[7] Ye Maosheng, Xu Shuangjie, Cao Tongyi. HVNet: Hybrid Voxel Network for LiDAR Based 3D Object Detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 1628-1637.

[8] Shi Shaoshuai, Guo Chaoxu, Jiang Li, et al. PV-RCNN: Point-voxel Feature Set Abstraction for 3D Object Detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 10526-10535.

[9] Chen Xiaozhi, Ma Huimin, Wan Ji, et al. Multi-view 3D Object Detection Network for Autonomous Driving[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 6526-6534.

[10] Engelcke M, Rao D, Wang D Z, et al. Vote3Deep: Fast Object Detection in 3D Point Clouds Using Efficient Convolutional Neural Networks[C]//2017 IEEE International Conference on Robotics and Automation (ICRA). Piscataway, NJ, USA: IEEE, 2017: 1355-1361.

[11] Li Bo. 3D Fully Convolutional Network for Vehicle Detection in Point Cloud[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ, USA: IEEE, 2017: 1513-1518.

[12] Qi C R, Liu Wei, Wu Chenxia, et al. Frustum PointNets for 3D Object Detection from RGB-D Data[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 918-927.

[13] Shi Shaoshuai, Wang Xiaogang, Li Hongsheng. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 770-779.

[14] Wang Zhixin, Jia Kui. Frustum ConvNet: Sliding Frustums to Aggregate Local Point-wise Features for Amodal 3D Object Detection[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ, USA: IEEE, 2019: 1742-1749.

[15] Yang Zetong, Sun Yanan, Liu Shu, et al. 3DSSD: Point-based 3D Single Stage Object Detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 11037-11045.

[16] Yang Zetong, Sun Yanan, Liu Shu, et al. STD: Sparse-to-dense 3D Object Detector for Point Cloud[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 1951-1960.

[17] Qi C R, Su Hao, Mo Kaichun, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 77-85.

[18] Qi C R, Yi Li, Su Hao, et al. PointNet++ : Deep Hierarchical Feature Learning on Point Sets in a Metric Space[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 5105-5114.

[19] Kuang Hongwu, Wang Bei, An Jianping, et al. Voxel-FPN: Multi-scale Voxel Feature Aggregation for 3D Object Detection from LIDAR Point Clouds[J]. Sensors, 2020, 20(3): 704.

[20] Lin T Y, Piotr Dollár, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 936-944.

[21] Geiger A, Lenz P, Urtasun R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2012: 3354-3361.

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

https://dc-china-simulation.researchcommons.org/journal/vol35/iss12/16
DOI: 10.16182/j.issn1004731x.joss.23-FZ0843E