J. Meas. Sci. Instrum., 2021, 12(2): 232-241 http://xuebao.nuc.edu.cn
DOI: 10.3969/j.issn.1674-8042.2021.02.013 jmsi@nuc.edu.cn

Adaptive multi-modal feature fusion for far and hard object detection

LI Yang[1,2], GE Hongwei[1,2]

(1. *Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi 214122, China;*
2. *School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China*)

Abstract: In order to solve difficult detection of far and hard objects due to the sparseness and insufficient semantic information of LiDAR point cloud, a 3D object detection network with multi-modal data adaptive fusion is proposed, which makes use of multi-neighborhood information of voxel and image information. Firstly, design an improved ResNet that maintains the structure information of far and hard objects in low-resolution feature maps, which is more suitable for detection task. Meanwhile, semantema of each image feature map is enhanced by semantic information from all subsequent feature maps. Secondly, extract multi-neighborhood context information with different receptive field sizes to make up for the defect of sparseness of point cloud which improves the ability of voxel features to represent the spatial structure and semantic information of objects. Finally, propose a multi-modal feature adaptive fusion strategy which uses learnable weights to express the contribution of different modal features to the detection task, and voxel attention further enhances the fused feature expression of effective target objects. The experimental results on the KITTI benchmark show that this method outperforms VoxelNet with remarkable margins, i.e. increasing the AP by 8.78% and 5.49% on medium and hard difficulty levels. Meanwhile, our method achieves greater detection performance compared with many mainstream multi-modal methods, i.e. outperforming the AP by 1% compared with that of MVX-Net on medium and hard difficulty levels.

Key words: 3D object detection; adaptive fusion; multi-modal data fusion; attention mechanism; multi-neighborhood features

0 Introduction

As an important step in the visual perception system, 3D object detection has been widely used in the fields of autonomous driving, robotics, virtual reality and augmented reality. LiDAR sensors are widely used in the field of autonomous driving and robotics due to their direct acquisition of three-dimensional (3D) structure information and accurate depth information of space targets. However, due to the shortcomings of sparseness and insufficient semantic information of point cloud data, it does not perform well in accurate far and hard object detection.

Most of the 3D object detection methods can be divided into the single sensor and multi sensors methods according to the modality of input data. The single sensor methods can be roughly categorized into grid-based method and PointNet-based method. Grid-based method transforms the point cloud into a regularly spaced grid to make full use of 2D or 3D convolutional networks, which can extract high-level representations of features from the grid. MV3D[1] uses a compact bird's-eye view to encode point clouds and preset multiple 3D anchor boxes to generate 3D bounding boxes. PIXOR[2] projects the point cloud to the bird's-eye view to obtain a dense and compact representation similar to the image, and then extracts the point cloud features by a 2D convolutional network. VoxelNet[3] is an end-to-end deep learning framework which uses the feature extractor layer to learn voxel features. Second[4] proposes an improved sparse convolution to replace 3D convolution which effectively reduces the amount of calculation and improves inference performance. PointPillar[5] uses vertical pillar to replace voxel units and uses 2D convolution to learn point cloud features, which improves the detection speed. Part-A2[6] is a new

Received date: 2021-02-26

Foundation items: National Youth Natural Science Foundation of China (No. 61806006); Innovation Program for Graduate of Jiangsu Province (No. KYLX160-781); Jiangsu University Superior Discipline Construction Project

Corresponding author: LI Yang (604509773@qq.com)

partial perception aggregation neural network, which uses partial perception modules and partial aggregation modules to improve target detection performance. PV-RCNN[7] combines the high efficiency of 3D convolution and the advantages of variable receptive field of PointNet-based method to improve the detection performance. Since the point cloud is sparse and uneven in nature, the sparse voxel grid brings a lot of redundant calculations and there is information loss in the process of voxelization and discretization. Instead, PointNet-based method directly processes the raw point cloud without information loss in the voxelization. PointNet[8] is an end-to-end deep neural network which directly learns the global features of the point cloud from the original point cloud. This method has good effects in 3D target recognition, instance segmentation and semantic segmentation. PointNet + +[9] improves PointNet again and can learn the local features of point clouds at different scales. PointRCNN[10] is a new two-stage detection framework. The first stage aims to generate 3D bounding boxes in a bottom-up scheme, and the second stage improves the 3D bounding boxes in standard coordinates. STD[11] is a new sparse to dense two-stage 3D target detection framework. The use of spherical anchor frames improves the target recall rate. The 3D IoU prediction branch improvement helps to align the classification confidence with the positioning confidence. VoteNet[12] proposes the Hough voting strategy to better group object features. A large number of point clouds will lead to high calculation and memory consumption. The performance of the above two methods will be worse when detecting objects from far distances due to the sparsenes and insufficient semantic information of the point cloud.

For the multi-sensors method, many state-of-the-art methods combine the data of multiple sensors to remedy the semantic loss of point clouds. MV3D[1] takes RGB-image, front-view and bird's-eye-view as input, and exploits a 3D region proposal networks (RPN) to generate 3D proposals. AVOD[13] uses a region proposal network to fuse multi-view features and generate target candidate regions. The second stage generates accurate object bounding boxes. MMF[14] uses correlated multi-task learning to fuse multi-modal features. MVX-Net[15] uses semantic image features to enhance the point cloud, and learns to fuse image and point cloud features at an early stage, which improves the performance of target detection. Frustum PointNets[16] first uses a mature 2D target detection algorithm to obtain the object proposal frame in the image. Then it uses the frustum to map proposal to the 3D space candidate area, and takes the PointNet-based models for target regression in the second stage.

However, the fusion methods such as MV3D[1] and AVOD[13] are too coarse because much background noise exists in the RoIs. Besides, these methods are difficult to detect small objects due to the fact that structure information of far and hard objects is seriously lost in high-level feature maps of deep networks. These methods simply resort to the feature pyramid network to acquire higher resolution feature map.

To overcome these shortcomings, based on the VoxelNet method, an improved ResNet[17] is firstly proposed to effectively maintain the structure information of far and small objects in high-level semantic feature maps. Besides, each image feature map is enhanced by semantic information from all subsequent feature maps, not just the subsequent layer in feature pyramid networks (FPN)[18]. After that, each point is enhanced by multi-level image semantic information in a point-wise manner. Secondly, multi-neighborhood context information of each voxel is obtained to solve the sparseness problem of point cloud. With multiple receptive fields with different sizes, the different context information will enhance the capability and robustness of voxel features to represent the spatial structure and semantic information of 3D objects contained in the voxel. Then, different modal features of point cloud and image are fused by this adaptive fusion strategy. Voxel attention further enhances the feature expression of effective target objects contained in the voxel and suppresses the expression of useless background features. Finally, the voxel features are sent to the convolutional network and region proposal network for target detection.

1 Proposed method

1.1 Overall network architecture

As shown in Fig. 1, the proposed network architecture is based on VoxelNet architecture marked by a dashed box. VoxelNet first voxelizes point cloud and then uses its voxel feature extractor (VFE) to extract voxel feature (VF). However, due to the sparseness of point cloud, each voxel lacks

enough points especially on far and hard objects, so the VF lacks sufficient semantic information to effectively represent 3D objects. Outside of the dashed box are our proposed component. The proposed network takes the point cloud and RGB image as input. To obtain the image features in point-wise manner, the improved ResNet is used as the image backbone. It is more suitable for far and hard object detection task to extract multi-level semantic features. The detailed structure information of far and hard objects is kept in high-level dilate-reslayer output feature map. Meanwhile, each feature map is enhanced by semantic information from all subsequent feature maps, not just the subsequent layer in FPN. After that, each point in the voxel is projected into the multi-level image feature maps from FPN. Bilinear interpolation is used to extract image features. Different image features are concatenated and then sent to a linear layer to form the final point-wise image features containing structural and semantic information, which can be used as prior knowledge to infer the presence of 3D object. Voxel-wise image feature (IF) can be obtained by point-wise max-pooling operation.

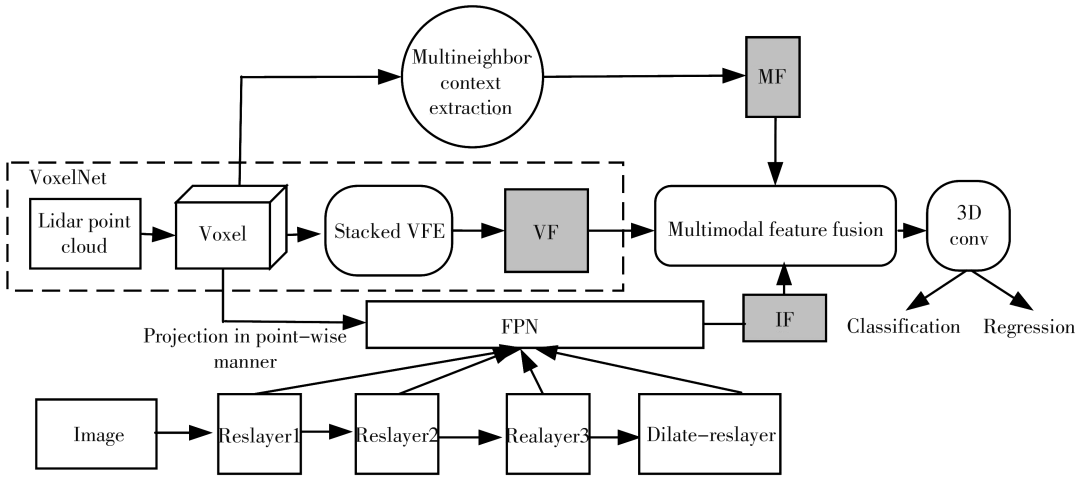


Fig. 1 Overall structure of proposed network

Secondly, the multi-neighborhood context information (MF) of each voxel is obtained to enhance the voxel features. At last, different modal features are fused adaptively by our fusion strategy and voxel attention further enhances the effective voxel features and suppresses the useless voxel features. Details of the improved ResNet, multi-neighborhood context information extraction and the proposed adaptive fusion technique are described in the following subsections.

1.2 Improved resNet and FPN

ResNet is usually used as the backbone network for classification tasks. In order to obtain abundant semantic information, a large enough receptive field is required, which will continuously reduce the resolution of the image. The detailed structure information of far and hard objects is severely lost, which is not suitable for target detection. Detection task requires not only semantic information for target classification, but also location information for target location regression.

In this work, to adapt RseNet to the detection task of far and hard objects and improve detection accuracy, the ResNet is improved by replacing the residual module with dilated residual module (dires-module), as shown in Fig. 2.

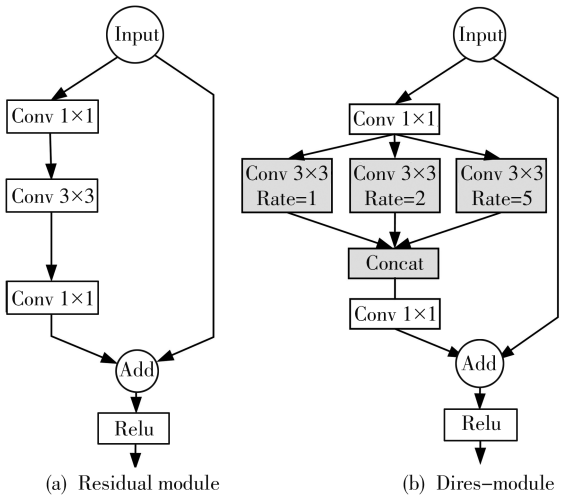


Fig. 2 Residual module and dires-module

As shown in Fig. 2(a), the residual module of ResNet is composed of three 2D convolutions with different kernel sizes. The original 2D convolution with kernel size of 3×3 in residual module is replaced with three parallel-distributed dilated convolutions. Each of them has the same kernel size of 3×3 but

different dilation rates of 1, 2 and 5. Dilated convolution can ensure a large enough receptive field without reducing the image resolution for deep network, so that the detailed structure information of far and hard objects will not be lost. At this time, a large resolution can provide detailed information for accurate bounding box regression, and large enough receptive field can provide semantic information for target classification. It effectively solves the problems of ResNet with high semantics and low resolution, which are not suitable for target detection. The features of three dilated convolutions are concatenated and fed into the convolution with kernel size of 1×1. Multiple dilation rates are set to provide different sizes of receptive fields to meet the detection requirements of objects at different distances. The receptive field with a low dilation rate is small and can effectively focus on short-distance information, and the receptive field with a high dilation rate is large and can focus on long-distance information. The combination of the three obtained multi-scale information is beneficial to the detection of near and far objects.

After obtaining the multi-level image features from improved ResNet, we hope to enhance the semantic information in low-level feature maps which helps to infer the presence of objects, so each layer of feature maps obtains semantic information from all subsequent feature maps, not just the subsequent layer in the FPN. As shown in Fig. 3, keep the reslayer1-reslayer3 of ResNet unchanged and replace reslayer4 with a dilate-reslayer. The dilate-reslayer consists of two stacked dires-modules. $P_1 - P_4$ represent the output feature maps from reslayer1 to dilate-reslayer, and $O_1 - O_4$ represent semantically enhanced output feature maps. The up arrow means the upsampling operation that doubles the size of the input image. The curve arrow means feature map concatenation. Note that all the upsampling operations output the same channel dimension. All the upsampled feature maps are as

$$\begin{aligned} P_{4-1} &= U(P_4),\\ P_{4-2} &= U(P_{4-1}),\\ P_{4-3} &= U(P_{4-2}),\\ P_{3-1} &= U(P_3),\\ P_{3-2} &= U(P_{3-1}),\\ P_{2-1} &= U(P_2), \end{aligned} \tag{1}$$

where U means the upsampling operations.

After upsampling from P_4, the spatial size of feature map P_{4-1} becomes twice as large and the spatial size of P_{4-2} gets forth as large and so on.

The O_1 is got by concatenating the P_{4-3}, P_{3-2}, P_{2-1} and P_1. The O_2, O_3 and O_4 can be obtained as follows

$$\begin{aligned} O_1 &= concat(P_{4-3}, P_{3-2}, P_{2-1}, P_1),\\ O_2 &= concat(P_{4-2}, P_{3-1}, P_2),\\ O_3 &= concat(P_{4-1}, P_3),\\ O_4 &= P_4. \end{aligned} \tag{2}$$

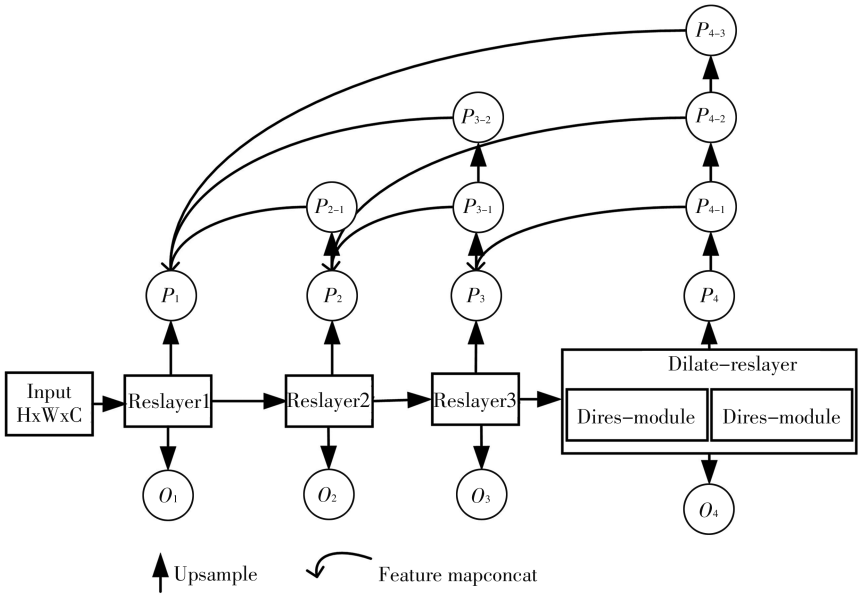


Fig. 3 Improved ResNet with FPN

After that, each output feature map $O_1 - O_4$ will get enough detailed structure information and semantic information which helps to infer the presence of target object. To obtain point-wise image

features, each point in the voxel is projected into the four image feature maps $O_1 - O_4$, respectively. Extract image features by using bilinear interpolation, then concatenate $O_1 - O_4$ features to obtain multi-level image features in point-wise manner.

1.3 Multi-neighborhood context information

VoxelNet randomly discards points during the voxelization process to ensure the same points number in each voxel. Furthermore, due to the sparseness of point cloud, the points number of far and hard objects contained in voxel is severely insufficient, so that the voxel feature cannot effectively characterize the structure of the object. To solve this problem, set up multiple receptive fields with different sizes to extract multiple neighborhoods context information and enhance the characterization ability and robustness of voxel features.

For the given voxel V, take its geometric center V_c as the sampling center and randomly sample K points in the neighbourhood of the sphere, the radius of sphere is not greater than u. The neighborhood set S of V_c is

$$S = \{[r_j; c_{V_c} - c_j] \mid \| c_{V_c} - c_j \| \leqslant u, \quad j = 1, \cdots, K\}, \tag{3}$$

where r_j is the reflection feature of point j; c_{V_c} and c_j are the world coordinates of V_c and point j respectively; K is the number of neighbor points of V_c. The coordinate offset and the point cloud feature are concatenated to indicate the local relative position of the point cloud feature, thereby effectively extracting context information. As shown in Fig. 4, after the sampling operation, we first get the input feature $(K, r+3)$, where r is the reflection feature and 3 is the coordinate offset.

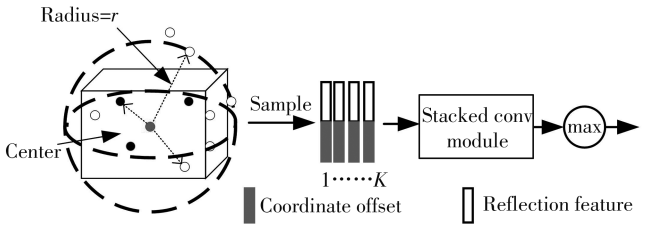


Fig. 4 Multi-neighborhood context information extraction

Next, each input feature is transformed through the stacked conv module into high dimensional feature space, where context information from features can be aggregated to encode the structure of objects contained in the voxel. The conv module is composed of a 2D convolution (stride of 1×1, kernel size of 1×1), a batch normalization (BN) and a rectified linear unit (ReLU). The transformed features belonging to the voxel can be aggregated by using element-wise max-pooling. At last, the MF of the voxel is obtained by concatenating the multiple context information.

1.4 Multi-modal feature fusion

VoxelNet takes a single-modal data as input and uses its VFE to aggregate the VF. Based on that, we try to adapt it to multi-modal data input to further improve its performance. In this work, an adaptive fusion strategy is proposed to fuse different modal features. Then further use the voxel attention to enhance the voxel feature expression of effective objects and suppress the feature expression of background objects. It is effective in improving the detection performance of far and hard objects in subsequent ablation experiments.

1.4.1 Adaptive fusion

Image data and point cloud data have different data characteristics and distributions, so we fuse them by learning the contribution degree of different modal features to the detection task through learnable weights. Useful features are assigned higher weights while useless features are assigned lower weights to achieve adaptive fusion of different modal features. As shown in upper part of Fig. 5, the VF, IF and MF are first concatenated and then sent to two linear layers. The sigmoid operation outputs the weight coefficient w_1 of each feature channel. VF, IF and MF are multiplied by the weight coefficient. After that, the weighted features are concatenated to output the fused voxel feature f.

1.4.2 Voxel attention

Voxel attention infers the presence of effective foreground target objects based on the voxel spatial position and fused voxel features f. The foreground target object features contained in the voxel are assigned higher weights to enhance voxel feature expression, and useless background object features contained in the voxel are assigned lower weights to suppress useless voxel feature expression. As show in lower part of Fig. 5, each fused voxel feature concatenates its world coordinates to provide accurate position information of objects. The concatenated features are sent to a linear layer. The sigmoid operation outputs the weight w_2. The final voxel

feature $v_1 - v_n$ are obtained by multiplying the w_2.

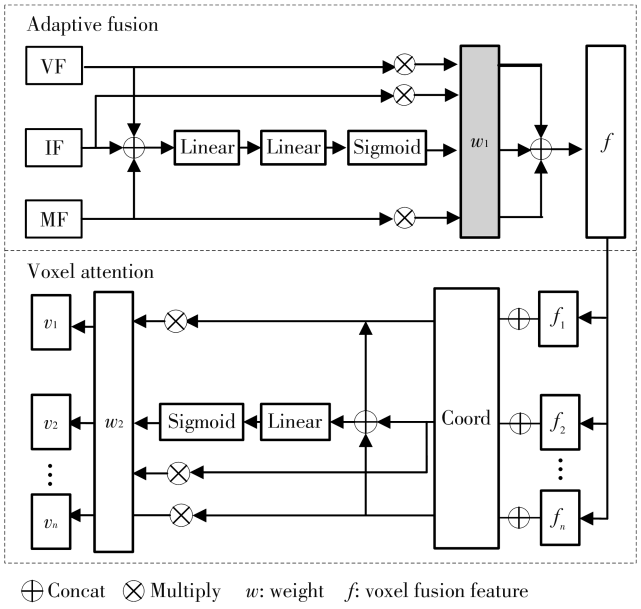


Fig. 5 Adaptive fusion and voxel attention

1.5 Training loss

The proposed framework is trained end-to-end with classification loss, regression loss and angle classification loss same with Second[4] as

$$L = \alpha_1 L_{\text{cls}} + \alpha_2 \sum_{r \in \{x,y,z,l,h,w,\theta\}} L_{smooth-L_1}(\Delta r, \Delta g) + \alpha_3 L_{\text{dir}}, \tag{4}$$

where the anchor classification loss L_{cls} is calculated by focal loss[19] with default hyper-parameters; smooth-L_1 loss is utilized for anchor box regression with the predicted residual Δr and the regression target Δg; the direction classification loss L_{dir} is calculated with cross entropy loss.

The overall training loss are the sum of these three losses with different loss weights α_1, α_2 and α_3, which are set to 1.0, 2.0 and 0.2, respectively.

2 Experiments

2.1 Experimental setup

2.1.1 Datasets

The proposed network is evaluated on the KITTI dataset[20] which is the one of the most popular datasets of 3D detection for autonomous driving. There are 7 481 training samples and 7 518 test samples which are divided into three difficulty levels: easy level, medium level and hard level based on the object size, occlusion and truncation. Since the label of the testing set cannot be obtained, we split the training set into train/validation sets to avoid the samples from the same sequence being included in both sets. After spliting the training samples, there are 3 712 samples in train set and 3 769 samples in validation set.

2.1.2 Metric

We adopt the average precision (AP) measured by 11 recall positions and 3 classes of mean average precision (mAP) as the metric to compare it with different methods. To prove the effect of proposed network on detecting far and hard objects, we mainly compare it with other methods at the medium and hard difficulty levels. During evaluation, we follow the official KITTI evaluation protocol: the IoU thresholds for class car, pedestrian and cyclist are 0.7, 0.5 and 0.5, respectively.

2.1.3 Network architecture

For the KITTI dataset, the detection range is from 0 m to 70.4 m for the X axis, from -40 m to 40 m for the Y axis and from -3 m to 1 m for the Z axis, which is voxelized with the voxel size of 0.05 m, 0.05 m and 0.1 m in each axis. So the voxel grids range is [41, 1 600, 1 408]. For image backbone, keep the reslayer1-reslayer3 of the ResNet-50 unchanged and replace the reslayer4 with a dilate-reslayer, which consists of two stacked dires-modules. The first 2D convolution with dires-module of stride 1×1 reduces the feature channel dimension from reslayer3 to 256, and each output feature channel dimension of each dilated convolution is set to 256, the last 2D convolution with stride 1×1 transforms the feature channel dimension from 768 to 1 024. The final output feature maps from improved ResNet are the four feature maps $P_1 - P_4$ with channel dimension of 256, 512, 1 024 and 1 024. Then use a 2D convolution to reduce the dimensionality of 256, 512, 1 024 and 1 024 to 128. For 3D convolution medium layers, employ 3D sparse convolution in Second[4] to speed up the inference time. We employ three phases of sparse convolution rather than two phases in Second. Each phase contains a sparse convolution to perform downsampling in each axis and two submanifold convolutions. The final output feature size is [5, 200, 176] reduced by 8 times and dimension of the feature is 256. For multi-neighborhood feature extraction, two different radius of 0.4 m and 0.8 m are set, the neighborhoods number K is set to 16, and the output feature dimension is set to 32.

2.1.4 Implementation details

The network is trained in an end-to-end manner by

the AdamW optimizer with an initial learning rate of 0.003, the betas are set to 0.95 and 0.99, respectively. Batch size is set to 1. The network is trained with 100 epochs.

For the data augmentation, since the point cloud and image multi-modal data are used at the same time, the data augmentation of the point cloud needs to be consistent with the image data augmentation, so individual ground-truth box augmentation strategy is not used. Random flipping, global rotation and global scaling are applied to the point cloud. The noise for global rotation is uniformly drawn from $-\pi/4$ to $\pi/4$ and the scaling factor is uniformly drawn from 0.95 to 1.05.

2.2 3D detection on KITTI dataset

2.2.1 Comparison with VoxelNet

The proposed network is based on VoxelNet, so we first compare its performance with VoxelNet on far and hard objects. Table 1 shows the results of comparison performance with VoxelNet. For the most important 3D object detection benchmark of the car class, our method outperforms VoxelNet with remarkable margins, i.e. increasing the AP by 8.78% and 5.49% on medium and hard difficulty levels. For the BEV detection of the car class, our method also achieves greater performance on medium and hard difficulty levels. As for the 3D detection and BEV detection of pedestrian and cyclist, our method outperforms VoxelNet significantly. The results prove the effectiveness of proposed method and overall network improves the detection performance of far and hard objects. The effectiveness of each component of the network will be explained in the ablation studies.

Table 1 Performance comparison with VoxelNet on far and hard objects with AP (%)

Detection	Method	Modality	Car		Pedestrian		Cyclist	
			Medium	Hard	Medium	Hard	Medium	Hard
3D	VoxelNet	LiDAR only	65.46	62.85	53.42	48.87	47.65	45.11
	Proposed method	RGB + LiDAR	74.24	68.34	56.97	51.40	51.50	46.42
	Improvement		+8.78	+5.49	+3.55	+2.53	+3.85	+1.32
BEV	VoxelNet	LiDAR only	84.81	78.57	61.05	56.98	52.18	50.49
	Proposed method	RGB + LiDAR	86.00	79.57	61.68	58.74	54.79	53.00
	Improvement		+1.19	+1.00	+0.63	+1.76	+2.61	+2.51

2.2.2 Comparison with mainstream multi-modal methods

To prove the superiority of our method, we compare it with multiple multi-modal methods on far and hard objects. As shown in Table 2, on the 3D objection detection benchmark of the car class, our method outperforms previous state-of-the-art methods on medium and hard difficulty levels.

Table 2 Performance comparison with mainstream multi-modal method on Car class with AP (%)

Method	Modality	3D AP(Car)	
		Medium	Hard
MV3D[1]	RGB + LiDAR	62.7	56.6
ContFuse[21]	RGB + LiDAR	66.2	64.0
F-PointNet[16]	RGB + LiDAR	70.9	63.7
MVX-Net[15]	RGB + LiDAR	73.3	67.4
Proposed method	RGB + LiDAR	74.2	68.3

In MVX-Net, there is information loss of far and hard object in high-level feature map from its VGG16 conv5 layer, while our method uses improved dilate-reslayer to keep the detailed structure information of far and hard objects in high-level feature maps which will be proved effectively in ablation studies.

2.2.3 Visualization of detection results

Part of the experimental results are projected onto the image for visualization. As shown in Fig. 6, the first row represents the groundtruth of the scene, including six cars in the vicinity, three cars in the distance and a very heavily blocked car.

Fig. 6 Visual display of predicted results

The second row represents the detection results of VoxelNet. VoxelNet successfully detected six nearby cars without occlusion and full of rich point cloud information. But VoxelNet missed the two farthest cars which were partially occluded and the point cloud was sparse (indicated by a cross). The third row is the detection results of proposed method. Our method successfully detected all the detection results of VoxelNet. Besides, it is worth noting that it successfully detected two partially occluded cars in the distance that VoxelNet missed (indicated by a tick). The visualization results show that the proposed method makes up for the defects of sparseness and insufficient information of point cloud by fusing multi-modal features and effectively improves performance in detecting far and hard objects.

2.3 Ablation studies

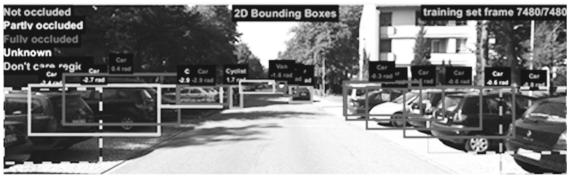
2.3.1 Effects of improved ResNet

The effects of improved ResNet was investigated by replacing it with unmodified ResNet and keeping all the modules unchanged including the reslayer4. All the ablation studies are evaluated 3 classes mAP of the car, pedestrian and cyclist on 3D car detection, which is more reliable and convincing. Table 3 shows that mAP drops about 1% and 0.5% on medium and hard difficulty levels when replacing improved ResNet, which validates that proposed ResNet can keep the detailed structure information of far and hard objects in high-level feature maps and is beneficial for inferring the presence of objects.

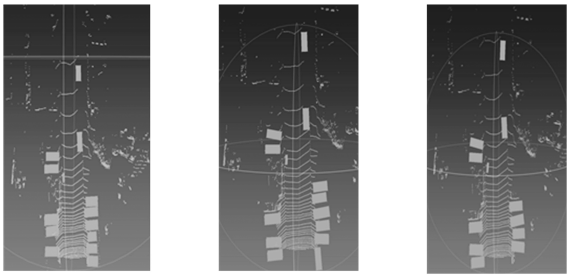
Table 3 Effects of proposed improved RseNet: mAP (%)

Method	3D mAP (3 class on car)	
	Medium	Hard
Unmodified ResNet	60.02	54.99
Improved ResNet	60.91	55.39

As shown in Fig. 7, Figs. 7(a) and (b) represent the ground truth bounding boxes in RGB image and point cloud. Figs. 7(c) and (d) are the predicted bounding boxes of improve ResNet and unmodified ResNet, respectively. Unmodified ResNet fails to detect the front right car which is far and seriously occluded, while our improved ResNet successfully detects all the ground truth bounding boxes. It proves that our method is robust and effective when detecting far and hard objects.



(a) Ground truth bounding boxes in RGB image



(b) Ground truth bounding boxes in point cloud

(c) Detection results of improved ResNet

(d) Detection resul of unmodified ResNet

Fig. 7 Visual display of predicted results of improved ResNet and unmodified ResNet

2.3.2 Effects of proposed fusion strategy and voxel attention

Table 4 validates the effectiveness of proposed adaptive fusion strategy and voxel attention. As shown in the first and second rows of Table 4, we first concatenate different modal features and use the voxel attention, the mAP improves 1.7% on medium difficulty level which proves the effectiveness of voxel attention that enhances the features expression of effective objects. To further improve the effects of adaptive fusion, we change the simple concatenation to our adaptive fusion strategy, the mAP improves by 2.60% and 1.19% greatly on medium and hard difficulty levels respectively, which validates the importance of different modal adaptive fusion. This benefits from that the adaptive fusion can tell importance of different modal features to the detection task by the learnable weights.

Table 4 Effects of proposed fusion strategy and voxel attention (%)

SC	VA	AF	3D mAP (3 class on Car)	
			Medium	Hard
√	×	×	56.61	54.18
√	√	×	58.31	54.20
×	√	√	60.91	55.39

(SC: Simply concat, VA: voxel attention, AF: adaptive fusion)

2.3.3 Effects of different components of network

As shown in Table 5, the importance of different components of proposed method was investigated. The first and second row show that the performance improves greatly by adding the improved RestNet which validates that the image information is effectively extracted and the point cloud features are

strengthened. The point cloud features are rich in semantic information, which is conducive to inferring the existence of objects. Furthermore, the mAP increases by a large margin on medium and hard difficulty levels by adding the multi-neighborhood context information. It proves that enough context information effectively makes up for the sparseness of point cloud and makes the voxel feature rich in the spatial structure and semantic information of 3D objects are useful for detecting far and hard objects.

Table 5 Effects of different components of network: mAP (%)

IR	MN	3D MAP (3 class on Car)	
		Medium	Hard
×	×	55.51	52.28
√	×	57.94	53.72
√	√	60.91	55.39

(IR: Improved ResNet, MN: Multi-neighborhood context)

3 Conclusions

In this paper, an adaptive multi-modal feature fusion for 3D object detection is proposed to solve the problem of sparseness and insufficient semantic information of single-modal point cloud and to improve the detection performance of far and hard objects. The improved RseNet is designed to extract point-wise multi-level image feature and maintain the detailed structure information of far and hard objects in high-level feature map simultaneously. Each multi-level feature map from improved RseNet is further enhanced by the semantic information from all subsequent feature maps, not just the subsequent layer in the FPN, which is beneficial to infer the presence of far and hard objects. Then the multi-neighborbood context information is extracted to make the voxel feature contained in the far and hard objects rich in spatial structure and semantic information of 3D objects, which is useful for detecting far and hard objects. The adaptive fusion strategy is proposed to fuse these different modal features according to their contribution to the detection task. The voxel attention further enhances the fused voxel features of effective target objects and supresses the invalid background objects features. All the proposed components are proved to be effective in ablation studies and the overall framework can significantly improve the detection performance of far and hard objects compared with VoxelNet and many mainstream multi-modal methods. Specifically, our method outperforms VoxelNet with remarkable margins, i. e. increasing the AP by 8.78% and 5.49% on medium and hard difficulty levels. Meanwhile, our method achieves greater detection performance compared with multi-modal method MVX-Net, the AP is increased by 1% on medium and hard difficulty levels.

References

[1] Chen X, Ma H, Wan J, et al. Multi-view 3D object detection network for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2017: 1907-1915.

[2] Yang B, Luo W, Urtasun R. PIXOR: real-time 3D object detection from point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018: 7652-7660.

[3] Zhou Y, Tuzel O. VoxelNet: end-to-end learning for point cloud based 3D object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018: 4490-4499.

[4] Yan Y, Mao Y, Li B. Second: sparsely embedded convolutional detection. Sensors, 2018, 18(10): 3337-3354.

[5] Lang A H, Vora S, Caesar H, et al. PointPillars: fast encoders for object detection from point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, 2019: 12697-12705.

[6] Shi S, Wang Z, Shi J, et al. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. doi:10.1109/TPAMI.2020.2977026.

[7] Shi S, Guo C, Jiang L, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, 2020: 10529-10538.

[8] Qi C R, Su H, Mo K, et al. PointNet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2017: 652-660.

[9] Qi C R, Yi L, Su H, et al. Pointnet++: deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems, 2017, 30: 5099-5108.

[10] Shi S, Wang X, Li H. PointRCNN: 3D object proposal generation and detection from point cloud. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, 2019: 770-779.

[11] Yang Z, Sun Y, Liu S, et al. STD: sparse-to-dense 3D object detector for point cloud. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, 2019: 1951-1960.

[12] Qi C R, Litany O, He K, et al. Deep hough voting for 3D object detection in point clouds. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, 2019: 9276-9285.
[13] Ku J, Mozifian M, Lee J, et al. Joint 3D proposal generation and object detection from aggregation. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, 2018: 1-8.
[14] Liang M, Yang B, Chen Y, et al. Multi-task multi-sensor fusion for 3D object detection. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, 2019: 7337-7345.
[15] Sindagi V A, Zhou Y, Tuzel O. MVX-Net: multimodal VoxelNet for 3D object detection. In: Proceedings of International Conference on Robotics and Automation (ICRA), Montreal, QC, 2019: 7276-7282.
[16] Qi C R, Liu W, Wu C, et al. Frustum pointNets for 3D object detection from RGB-D data. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018: 918-927.
[17] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016: 770-778.
[18] Lin T, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017: 936-944.
[19] Lin T, Goyal P, Girshick R, et al. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.
[20] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, 2012: 3354-3361.
[21] Liang M, Yang B, Wang S, et al. Deep continuous fusion for multi-sensor 3D object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018: 641-656.

自适应性多模态特征融合的远小困难目标检测

李　阳[1,2]，葛洪伟[1,2]

（1. 江南大学 江苏省模式识别与计算智能实验室，江苏 无锡 214122；
2. 江南大学 人工智能与计算机学院，江苏 无锡 214122）

摘　要： 为了解决由 LiDAR 点云稀疏性和语义信息不足造成的远小困难物体检测困难的问题，提出了一种多模态数据自适应性融合的 3D 目标检测网络，充分融合了体素的多邻域上下文信息和图片多层语义信息。首先，设计了一种更适用于检测任务的改进残差网络，提取图片多层语义特征的同时，在低分辨率特征图中有效保留了远小物体的结构细节信息。每个特征图进一步通过来自所有后续特征图的语义信息进行语义增强。其次，提取具有不同感受野大小的多邻域上下文信息，弥补远小物体点云信息不足的缺陷，加强体素特征的结构信息和语义信息，以提高体素特征对物体空间结构和语义信息的表征能力及特征鲁棒性。最后，提出了一种多模态特征自适应融合策略，通过可学习权重，根据不同模态特征对检测任务的贡献程度进行自适应性融合。此外，体素注意力根据融合特征进一步加强有效目标对象的特征表达。在 KITTI 数据集上的实验结果表明，本方法以明显的优势优于 VoxelNet，即在中等难度和困难难度下 AP 分别提高 8.78% 和 5.49%。同时，与许多主流的多模态方法相比，本方法在远小困难物体的检测性能上具有更高的检测性能，即在中等和困难难度级别上，AP 的性能比 MVX-Net AP 均高出 1%。

关键词： 3D 目标检测；自适应性融合；多模态数据融合；注意力机制；多邻域特征

引用格式： LI Yang, GE Hongwei. Adaptive multi-modal feature fusion for far and hard object detection. Journal of Measurement Science and Instrumentation, 2021, 12(2): 232-241. DOI: 10.3969/j.issn.1674-8042.2021.02.013