International Journal of Trend in Research and Development, Volume 6(5), ISSN: 2394-9333
www.ijtrd.com

3D Object Detection based on AVOD algorithm

[1.*]Huailong Yi, [2]Zhuangzhuang Mao and [3]Mengchao Liu,
[1,2,3]School of Mechanical and Power Engineering, Henan Polytechnic University, Jiaozuo, China

***Abstract*—**3D objectdetection plays an important role in a large number of real-world applications. 3D object detection has achieved high accuracy and efficiency, but the small object detection is still a challenge. In view of the current low detection accuracy of the small objects, the paper studies the multimodal fusion AVOD model to detect cars, pedestrians and cyclists, and fine-tunes the model. The small object detection method based on the skip feature pyramid model is introduced to fuse the detailed information of the multi-layer high-level semantic feature information and the low-level feature map, the object detection accuracy of the model is further improved.The experimental results on the KITTI datasets show that the proposed approach obtains significant improvements.

Keywords: 3D *Object Detection; AVOD; feature pyramid;multi-modal fusion; deep convolutional neural network.*

I. INTRODUCTION

In recent years, with the emergence of 2D object detection algorithms such as Faster-RCNN[1], SSD[2], and Yolo[3], deep Convolutional Neural Networks have made significant progress in the study of 2D object detection in complex scenes. However, one of the challenges of systems based on 2D images is that it is difficult for cameras to provide accurate 3D location information. Therefore, based on the above 2D object detection algorithm, the research on related variants of object detection algorithm is also rapidly advancing. The precision of 3D object detection is still a problem to be solved. The most important thing of 3D object detection is how to use the 3D information to express the depth information of the object in the estimation task. The key challenge is that the point cloud data captured by LIDAR sensors are sparse and irregular. The problem of small objects detection brings us great challenges.

The abundant environmental texture information is provided by camera images. However, objects can be occluded, and the scale of a single object can vary greatly on the camera imaging plane. LIDAR point clouds provide information about the depth and reflectivity of the environment. For 3D environment inference, the Bird's Eye View(BEV) of LIDAR point clouds may be a better representation method. In order to facilitate the input of neural network, some methods transform 3D point clouds data into voxel format to achieve object detection .VoxelNet [4] takes the original point cloud as input, divides the space into voxels, and converts the points in each voxel into vectors representing shape information. However, it is very expensive in computation for the high dimensionality of point clouds. F-PointNet [5]firstly generates a frustum sequence for each region proposal under the proposal of a given 2D region in the RGB image, and then uses the frustum to group local points, so as to continuously estimate the 3D box in 3D space from end to end. However, 3D object detection is restricted by 2D object detection. Many approaches [6,7] project the point cloud into the BEV feature map and apply 2D convolution neural network (CNN) to these feature maps for 3D object detection, which makes more effective use of 3D data. The two-stage detection framework of MV3D [6] includes RPN and detection. The network uses BEV, Front View(FV) and RGB images as input. Firstly, 3D object proposals are generated from the BEV, and then they are projected to three views: BEV, FV from LIDAR and RGB images. A deep fusion network is used to combine the regional features obtained from each view through ROI pooling. Fusion features are used to predict object classes jointly and orientated 3D box regression。 Because MV3D only uses BEV to propose candidate regions, it performs well on large objects such as cars. However, it will not have good performance in detecting the small objects such as pedestrians and cyclists. AVOD [7] further improves MV3D, and uses BEV and RGB images to propose regional candidates. The model can be fused to generate region proposals for two-stage object detector. The regional multimodal features of each scheme can also be fused.

Inspired by AVOD, the paper further studies the regional multi-modal feature fusion method for each scheme. AVOD uses two identical feature extractors, one for RGB image input and one for LIDAR BEV input. In the feature extraction part, the final feature maps obtained by encoder and decoder are characterized by high resolution and representativeness, and shared by RPN and the second stage detection network. On this basis, we use the skip feature pyramid model to fuse high-level and low-level feature map information, which is more efficient for small object detection.

II.AVOD ARCHITECTURE

AVOD takes the RGB images and the BEV as input. Two identical feature extractors are used to extract the features of the RGB images and the BEV respectively. These feature maps are passed to the second stage detection network for dimension refinement, direction estimation and classification. The network structure is shown in Figure 1.

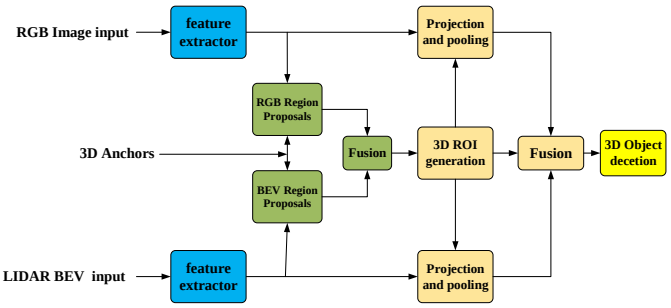

Figure 1: AVOD Network Architecture.

A. The skip feature pyramid model

The two identical feature extractors proposed by AVOD are shown in Figure 2, which are composed of an encoder and a decoder. The encoder has made some modifications to VGG-16, mainly reducing the number of channels by half, and cutting the network in Conv4 layer. Therefore, the encoder takes an $M \times N \times D$ image or a bird's eye view as input and outputs $(M/8) \times (N/8) \times D$ feature map. In the KITTI datasets, pedestrians are usually 0.8×0.6 m, and occupy 8×6 pixels in the BEV (resolution is 0.1m). After 8 times down-sampling by the encoder, only one pixel is occupied in the output feature map. A bottom-up decoder is designed to restore the sample

from the feature map of the encoder to the original input size, and a conv-transpose is used to concatenate the associated feature map of the two encoder, then the two coders are fused by a 3×3 convolution.The final feature maps obtained by encoder and decoder are characterized by high resolution and representativeness, and shared by RPN and the second stage detection network.

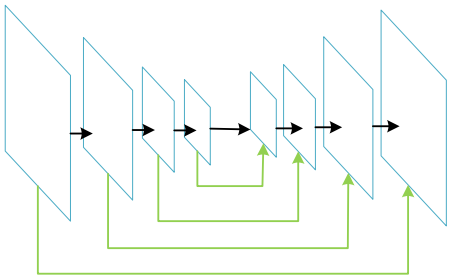

Figure 2: Encoder and Decoder of the AVOD.

In the paper, the decoder is further improved. We adopt skip connection method to sample high-level feature maps by choosing deconvolution with different step sizes, and use the pixel-by-pixel summation method to fuse the information between non-adjacent feature maps. The skip feature pyramid model is shown in Figure 3. In deep convolution network, the deepest feature map of the network contains the most abstract feature information. Therefore, using the proposed skip feature pyramid model to fuse the information between different high-level and low-level feature maps can not only effectively utilize the scale information between different feature layers, but also fuse the detail information between high-level feature maps and low-level feature maps. By choosing the high-level feature map in the basic network, each high-level feature map is convoluted using convolution kernels with the size of 3×3 channels, which are 32, 64, 128 and 256, respectively. The purpose of this method is to change the number of channels of feature maps of different feature layers into the same number for fusion calculation. After unifying the number of channels in each layer, the deconvolution operation with a size of 2×2 is used to adopt the feature map of adjacent feature layers. After the upper sampling, the different feature layers become the same size. Then the non-adjacent feature maps are sampled by deconvolution with a size of 4×4.

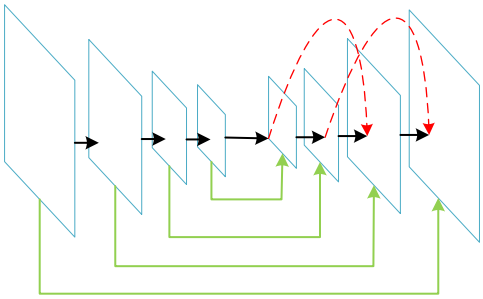

Figure 3: Encoder and Decoder of the changed AVOD.

B. Loss function

Similar to RPN, we use two smooth L1 loss combinations for boundary box and direction vector regression tasks, and cross-entropy loss for classification tasks. During training, the positive/negative ROI are determined based on the IoU overlap of in the BEV boxes. For cars, the anchors are assigned to ground-truth objects using an intersection-over-union (IoU) threshold of 0.5 and are assigned to the background (negative) if their IoU are less than 0.45. For pedestrians and cyclists, we use values of 0.45 for the nonmatching threshold and 0.5 for the matching threshold. In order to eliminate overlapping detection, NMS with a threshold of 0.01 is used.

$$\text{cross\_entropy}(p,y)=\begin{cases}-\log(p) & \text{if } y=1\\ -\log(1-p) & \text{otherwise,}\end{cases}$$

$$\text{smooth}_{L1}(x)=\begin{cases}0.5x^2 & \text{if } |x|<1\\ |x|-0.5 & \text{otherwise.}\end{cases}$$

III. EXPERIMENTS AND RESULTS

A. Implementation Details

The point cloud is cropped at [−40,40] × [0,70] meters to contain points within the field of view of the camera. For bird's eye view, the discretization resolution is set to 0.1m, therefore the bird's eye view input has size of 704×800. Since KITTI uses a 64-beam Velodyne laser scanner, we can obtain a 64×512 map for the front view points. We set the height range to [0,2.5] meters along the Z axis. The network is trained in an end-to-end way. We use SGD to train the network. The total number of iterations is 120,000. The initial learning rate is 0.0001.Our computation environment for inference included a Core-i5 8400 CPU, 16 GB of DDR4 memory and a ZOTAC GTX 1070 8G Ti GPU.

B. Evaluation Metrics

We evaluate our approach on the challenging KITTI dataset [8] which provides 7481 images for training and 7518 images for testing. Following the standard KITTI setup, we use the Average Precision (AP) metric for the object detection task. We evaluate three levels of difficulty: easy, moderate, and hard. In our calculations, we set 3D IoU thresholds of 0.5 and 0.7, respectively.

C. Evaluation Result

Loss analysis: In order to evaluate the convergence of the model, this experiment visualizes the classification loss of vehicles and pedestrians, the regression loss and the total training loss of the model. As shown in Figure 4, it can be seen from the curve that the model has good convergence.

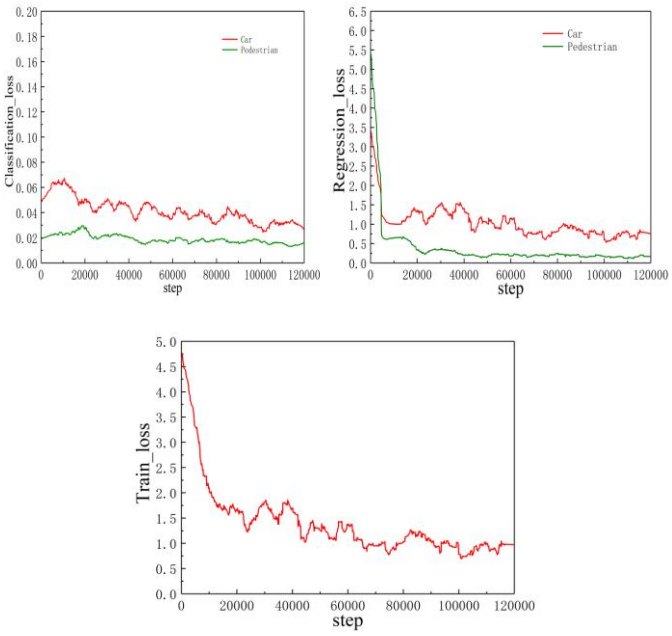

Figure 4: Loss curve

***3D object Detectionprecision*:** In this paper, the decoder part of AVOD is further improved. The skip feature pyramid model is shown in Figure 3. We evaluate the categories of cars,

pedestrians, and cyclists on the KITTI dataset. The thresholds are taken as 0.5 and 0.7, respectively. The results are shown in Table 1, Table 2 and Table 3. As can be seen from the above table, the skip feature pyramid model has higher detection precision for cars, pedestrians and cyclists than the feature pyramid of the original decoder. Especially for pedestrians and cyclists, the small object detection be improved greatly.We also visualized the average precision curve for car, pedestrian, and cyclist at the threshold of 0.7, as shown in Figure 5. The FPN in the table represents the original feature pyramid network, and the SFPN represents the skip feature pyramid network.

Table 1: 3D detection performance: Average Precision ($AP_{car\text{-}3D}$) (in %) of 3D boxes on KITTI test set

Method	IoU=0.5			IoU=0.7		
	Easy	Mod.	Hard	Easy	Mod.	Hard
AVOD+FPN	90.18	88.86	88.16	81.94	71.88	66.38
AVOD+SFPN	90.42	89.08	88.26	82.42	73.02	67.12

Table 2: 3D detection performance: Average Precision ($AP_{pedestrian\text{-}3D}$) (in %) of 3D boxes on KITTI test set

Method	IoU=0.5			IoU=0.7		
	Easy	Mod.	Hard	Easy	Mod.	Hard
AVOD+FPN	59.74	52.71	45.93	48.42	42.50	36.71
AVOD+SFPN	63.01	55.02	47.82	51.54	44.51	37.97

Table 3: 3D detection performance: Average Precision ($AP_{cyclist\text{-}3D}$) (in %) of 3D boxes on KITTI test set

Method	IoU=0.5			IoU=0.7		
	Easy	Mod.	Hard	Easy	Mod.	Hard
AVOD+FPN	55.54	37.97	31.55	52.79	30.97	30.00
AVOD+SFPN	57.15	38.17	32.22	55.48	32.19	31.42

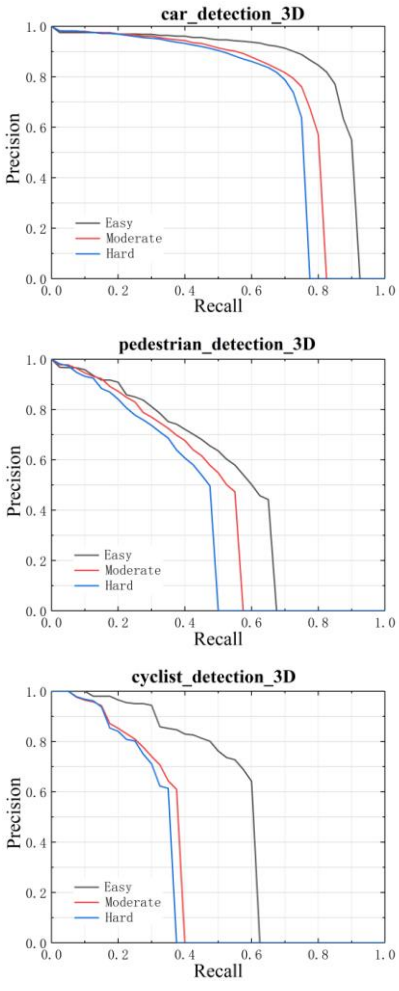

Figure 5: Precision-Recall curve

Qualitative Results: Figure 6 shows some qualitative results of our approach. Our method can handle different scenes. It is robust to small object in different distance. And when the scene is crowded, our method still performs well in most cases. For pedestrian and cyclist detection, there are omissions and errors. Because the number of pedestrian and cyclist samples on the KITTI data set is low. The ground truths are in red. The prediction boxes for the three categories are shown by green, blue and yellow respectively.

Figure 6: Qualitative results of 3D object detection on KITTI benchmark

CONCLUSION

In this paper, AVOD model is used to realize 3D object detection. The small object detection method based on the skip feature pyramid model is introduced to fuse the detailed information of the multi-layer high-level semantic feature information and the low-level feature map, the object detection accuracy of the model is further improved. We adopt skip connection method to sample high-level feature maps by choosing deconvolution with different step sizes, and use the pixel-by-pixel summation method to fuse the information between non-adjacent feature maps.Our approach can handle different scenes. It is robust to small object in different distance.

References

[1] Q. Ren, K. M. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,"IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1137–1149, 2017

[2] W. Liu, D. Anguelov, D. Erhan, and C. Szegedy, et al.,"SSD: Single Shot MultiBox Detector," Lecture Notes in Computer Science, pp. 21–37, 2016

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi,"You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 779–788, 2016

[4] Y. Zhou and O. Tuzel, "VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection," in Proceedings of the

IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4490–4499, 2018

[5] C. R. Qi, W. Liu, and C. X. Wu, et al., "Frustum PointNets for 3D Object Detection from RGB-D Data," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 918–927, 2018

[6] X. Z. Chen, H. M. Ma, and J. Wan, et al., "Multi-view 3D object detection network for autonomous driving," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, pp. 6526–6534, 2017

[7] J. S. Ku, S. M. Mozifian, and J. Lee, et al., "Joint 3D Proposal Generation and Object Detection from View Aggregation," IEEE International Conference on Intelligent Robots and Systems, pp. 5750–5757, 2018

[8] A. Geiger, P Lenz, P. Girshick, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3354–3361, 2012