3D Object Detection with Differential Point Clouds

Subjects: Computer Science, Artificial Intelligence

Contributor: Guangjie Han, Yintian Zhu, Lyuchao Liao, Huiwen Yao, Zhaolin Zhao, Qi Zheng

3D object detection based on point clouds has many applications in natural scenes, especially in autonomous driving. Point cloud data provide reliable geometric and depth information.

Keywords: hybrid attention ; autonomous driving ; point cloud

1. Introduction

3D object detection based on point clouds has many applications in natural scenes, especially in autonomous driving. Point cloud data provide reliable geometric and depth information. However, point clouds are disordered, sparse, and unevenly distributed, increasing the difficulty of object detection [1].

Currently, existing object detection methods mainly include image-based, point cloud-based, and multi-sensor methods [2]. In comparing them, image-based methods lack depth and 3D structure information, making it challenging to identify and locate 3D objects accurately in 3D space. Therefore, plans based on image information tend to be less effective than point clouds [3][4][5]. GS has proposed fusing point cloud and image data for object detection [6]. Subsequently, the classic methods MV3D, PC-CNN [7], AVOD [8], PointPainting [9], etc., have been proposed. However, although these fusion methods can integrate the characteristics of point clouds and images to a certain extent for recognition, the vast amount of calculation involved and the complex network has brought considerable challenges to this field. Thus, point cloud-based methods are the main methods for autonomous driving. The method based on the point cloud has developed rapidly in the last few years, and many classic methods have been proposed, including Pointnet [10], Pointnet++ [11], VoxelNet [12], SE-SSD [13], etc.

Early works usually convert raw point clouds into regular intermediate representations, including projecting 3D point cloud data from bird's eye or frontal views into 2D images or dense 3D voxels. However, using voxel conversion to improve efficiency can lead to a lack of critical information, resulting in false and missed detection. PointPillars [14] encode point clouds with Pillar coding, which achieves extremely fast detection speed. However, it loses many important foreground points simultaneously, making the effect of detail processing not ideal. There have been a lot of missed and false detections in PointPillars. To solve this critical problem, TANet [15] enhances the local characteristics of the voxel by introducing an attention mechanism. However, due to the information loss during voxel conversion, it is impossible to avoid the occurrence of false and missed detection. In DA-PointRCNN [16], the density sampling method can pay better attention to where the clouds are sparse and improve missed detection. However, false detection exists due to ignoring the importance of feature information.

2. Voxel-Based Methods

In point cloud-based methods, converting the raw point cloud into a regular voxel grid and extracting local features for object detection has attracted much attention. the The voxel concept was first proposed with VoxelNet, in which the point cloud is divided into voxels by block and detected by extracting local features from each voxel. However, even this requires considerable computation. SECOND [17] adds a sparse convolution operation based on VoxelNet to speed up calculation. PointPillars directly converts point clouds into fake images, avoiding the time-consuming convolution calculation.

According to their different detection stages, the existing voxel detectors can be roughly divided into single-stage detectors and two-stage detectors. While these methods are efficient and straightforward, due to the reduction of spatial resolution and insufficient structural information their detection performance is significantly affected when the point cloud is relatively sparse. Thus, SA-SSD [18] supplements the utilization of structural information by adding auxiliary networks. HVNet [19] offers a hybrid voxel network that refines the projected and aggregated feature maps from multiple scales to improve detection performance. CIA-SSD [20] introduces a network incorporating IOU-aware confidence correction to

extract spatially informative features of detected objects. In comparison, two-stage detectors can achieve higher performance at the cost of higher computation and storage. Part-A2 [21] proposes a two-stage detector consisting of part perception and aggregation modules, which is better able to utilize the location information of detected objects.

In general, detection methods based on voxel detection can achieve better detection effects and higher efficiency to a large extent. However, voxelizing the point cloud inevitably causes information loss. Later research work has made up for the loss and distortion caused by the point cloud data processing stage by continuously introducing complex module designs, which has made up for this defect to a certain extent; however, this has a great impact on detection efficiency. Therefore, using voxelization to process point cloud data has certain limitations.

3. Point-Based Methods

Unlike voxel-based detection methods, point-based methods directly process the disordered and cluttered point cloud. Thus approach obtains features point-by-point in order to predict each point. The point cloud itself contains very rich physical structure information. Therefore, a point-wise processing network was first proposed in the form of PointNet. This network directly takes the original point cloud as input, guaranteeing no loss of physical information from the original point cloud. Subsequently, PointNet++ improved PointNet to improve the detection efficiency of the network and further optimize the network structure. Most of the subsequent point-based methods have used this network and its variants to point cloud for processing. PointRCNN [22] utilizes PointNet++ to extract features from raw point clouds and a Region Prediction Network (RPN) to generate prediction boxes. 3DSSD [23] introduces a 3D single-stage detection network which uses Euclidean space to achieve feature sampling for far points. PointGNN [24] adds a graph neural network to the framework of 3D object detection, effectively improving recognition accuracy. Proposal Contrast [25] proposed a new unsupervised point cloud pre-training framework to achieve better detection results. Proficient Teachers [26] introduces a new 3D SSL framework that provides better results and removes the necessity of using confidence-based thresholds to filter pseudo-labels.

Point-based detection methods directly process the raw point cloud and effectively utilize the physical information of the point cloud itself. However, the huge amount of data inevitably takes up a lot of time and computing resources. Therefore, improving the efficiency of point-based detection is a bottleneck for this method.

4. Hybrid Attention Regions with CNN Features (HA-RCNN)

Unlike voxel-based methods, point-based methods need to perform point-wise detection, and as such need to pay more attention to foreground points (i.e., cars, pedestrians, etc.). However, most current point-based object detection frameworks usually adopt downsampling methods, such as random sampling [27] or farthest point sampling. Although these sampling methods can improve computational efficiency, the essential foreground points are ignored. Therefore, in this research, researchers aim to train a point-based model to better retain the information of foreground points and efficiently detect multiple types of objects at one time. Based on this, researchers propose an efficient point cloud-based object detection algorithm.

As shown in **Figure 1**a, the proposed model framework mainly consists of three parts: Hybrid Sampling (HS), a Hybrid Attention Mechanism (HA), and Foreground Point Segmentation. First, the input original point cloud is processed through hybrid sampling, with as many foreground points retained as possible. Then, the point-wise features are generated by the HA module and focused. Subsequently, the foreground segmentation network is used to segment the foreground points and generate prediction boxes. Finally, 3DNMS is used to filter the prediction box and the refinement module retains the final boxes. In **Figure 1**b, each sampled point cloud input is extracted pointwise and then focused in the attention layer. Finally, the generated original pointwise features and the pointwise features developed by the attention layer are spliced together.

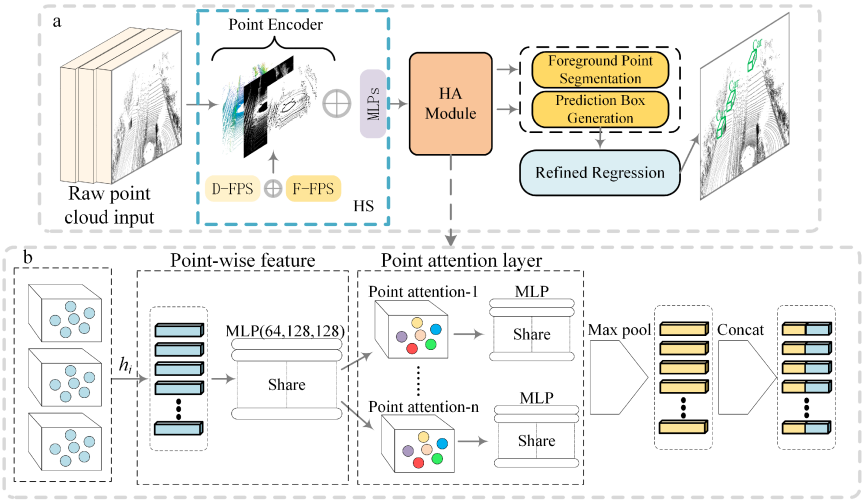

Figure 1. HA-RCNN model frame diagram (**a**: Overall frame, **b**: HA module refinement).

References

1. Huang, T.; Liu, Z.; Chen, X.; Bai, X. Epnet: Enhancing point features with image semantics for 3d object detection. In European Conference on Computer Vision, Proceedings of the Computer Vision ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer: Cham, Switzerland, 2020; Volume 12360, pp. 35–52.
2. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.
3. Qin, Z.; Wang, J.; Lu, Y. Monogrnet: A geometric reasoning network for monocular 3d object localization. In Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8851–8858.
4. Li, B.; Ouyang, W.; Sheng, L.; Zeng, X.; Wang, X. Gs3d: An efficient 3d object detection framework for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1019–1028.
5. Liu, Z.; Zhou, D.; Lu, F.; Fang, J.; Zhang, L. Autoshape: Real-time shape-aware monocular 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15641–15650.
6. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In European Conference on Computer Vision, Proceedings of the Computer Vision ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; Volume 8695, pp. 345–360.
7. Du, X.; Ang, M.H.; Karaman, S.; Rus, D. A general pipeline for 3d detection of vehicles. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 3194–3200.
8. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. Joint 3d proposal generation and object detection from view aggregation. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Madrid, Spain, 1–5 October 2018; pp. 1–8.
9. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. Pointpainting: Sequential fusion for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 4604–4612.
10. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
11. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inf. Process. Syst. 2017, 30, 5099–5108.

12. 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 (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4490–4499.
13. Zheng, W.; Tang, W.; Jiang, L.; Fu, C.W. SE-SSD: Self-ensembling single-stage object detector from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 14494–14503.
14. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 12697–12705.
15. Liu, Z.; Zhao, X.; Huang, T.; Hu, R.; Zhou, Y.; Bai, X. Tanet: Robust 3d object detection from point clouds with triple attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11677–11684.
16. Li, J.; Hu, Y. A Density-Aware PointRCNN for 3D Object Detection in Point Clouds. arXiv 2020, arXiv:2009.05307.
17. Yan, Y.; Mao, Y.; Li, B. SECOND: Sparsely Embedded Convolutional Detection. Sensors 2018, 18, 3337.
18. Du, L.; Ye, X.; Tan, X.; Feng, J.; Xu, Z.; Ding, E.; Wen, S. Associate-3Ddet: Perceptual-to-conceptual association for 3D point cloud object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 13329–13338.
19. He, C.; Zeng, H.; Huang, J.; Hua, X.S.; Zhang, L. Structure aware single-stage 3d object detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11873–11882.
20. Ye, M.; Xu, S.; Cao, T. Hvnet: Hybrid voxel network for lidar based 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 1631–1640.
21. Shi, S.; Wang, Z.; Shi, J.; Wang, X.; Li, H. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 43, 2647–2664.
22. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3d object proposal generation and detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 770–779.
23. Yang, Z.; Sun, Y.; Liu, S.; Jia, J. 3DSSD: Point-Based 3D Single Stage Object Detector. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, WA, USA, 14–19 June 2020; pp. 11037–11045.
24. Shi, W.; Rajkumar, R. Point-GNN: Graph Neural Network for 3D Object Detection in a Point Cloud. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Seattle, WA, USA, 14–19 June 2020; pp. 1708–1716.
25. Yin, J.; Zhou, D.; Zhang, L.; Fang, J.; Xu, C.Z.; Shen, J.; Wang, W. Proposalcontrast: Unsupervised pre-training for lidar-based 3D object detection. In European Conference on Computer Vision, Proceedings of the Computer Vision ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; Volume 13699, pp. 17–33.
26. Yin, J.; Fang, J.; Zhou, D.; Zhang, L.; Xu, C.Z.; Shen, J.; Wang, W. Semi-supervised 3D object detection with proficient teachers. In European Conference on Computer Vision, Proceedings of the Computer Vision ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; Volume 13698, pp. 727–743.
27. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Seattle, WA, USA, 14–19 June 2020; pp. 11105–11114.