3D Object Detection and tracking in Point Cloud Flow

BMVC 2019 Submission # ??

Abstract

Recent approaches for 3D object detection has made great progresses due to the evolution in deep learning. However, previous works are mostly based on single frame point cloud or image, information between point cloud frames is almost not utilized. In this paper, we try to leverage the temporal information in point cloud flow and explore 3D object detection and tracking based on 3D data flow. Towards this goal, we set up a ConvNet architecture that can associate multi-frame image and LIDAR data to produce accurate 3D detection boxes and trajectories. Notably, a correlation module is introduced to capture object co-occurrences across time, and a multi-task objective for frame-based object detection and across-frame track regression is used, therefore, the network can performs detection and tracking simultaneously. Our proposed architecture is shown to produce competitive results on the KITTI Object tracking datasets. Code and models will be available soon.

1 Introduction

A lot of attention has been paid to 3D object detection because of the rapid development of autonomous driving industry in recent years. Although object detection in images has made tremendous progress due to the emergence of deep learning [5, 6, 11] and their region based descendants [2, 4, 10], extending 2D approaches to 3D scene is extremely hard. This is mainly because of the curse of dimensionality and the sparsity of 3D data.

In spite of the difficulty in 3D object detection, many works have been carried out. Recent approaches are usually done in three fronts: image based, point cloud based and fusion of image and point cloud. These methods have achieved remarkable performance but are limited to single frame input, applying existing detection networks on individual frames will loss the consistency and difference between frame and introduce unaffordable computational cost for most applications.

Compare to single point cloud frame, point cloud flow data is more natural and straightforward for most situation. Thus Fast and accurate point cloud flow object detection is crucial for autonomous driving. Similar to extend 2D object detection methods to 3D situation, we also can extend video object detection approaches to 3D point cloud flow object detection.

Most modern computer vision approaches to video object detection require flow estimation, which is a fundamental task in video analysis. For example, a series work in [12, 13] associate vision feature and optical flow to build an accurate and end-to-end learning framework for video object detection. However, applying video object detection framework to

autonomous driving is hard, because video object detection suffers from motion blur and partial occlusion, which are conventional in driving scenarios. While another choice is to utilize LIDAR data, since point cloud provides an accurate spatial representation of the world allowing for precise positioning of objects of interest, motion blur and occlusion problem can be avoided in 3D object detection. However, LIDAR does not capture appearance well when compared to the richness of images, moreover, accurate 3D scene flow estimation from point clouds is tremendously tough. Although some approaches such as [1, 2] have been presented to learn 3D motion field in the world, it is hard to implementation in large scenes. Thus the challenge is, can we propose an approach that accurately do 3D object detection in point cloud flow without 3D scene flow?

Inspired by [3], we transform AVOD [7] structure into a two stream network embedding with a correlation module, named Bi-AVOD, which takes two adjacent key frames as input and predicts location and orientation of object as well as their local displacement. Noting that Bi-AVOD is an aggregate view object detection architecture capable of fusing different features in image and point cloud, thus its input includes two adjacent images in front view and two adjacent BEV (bird eye view) from LIDAR data. While the correlation module compute convolutional cross-correlation between the feature responses of adjacent key frames to estimate displacement of the same objects. With local displacement and object orientation, object location in intermediate frames can be calculated by interpolation. Moreover, detections can be linked between frames with the help of local displacement and multiple object tracking can be performed through *tracking by detection* [8].

In summary, our contributions are threefold: *(i)* we set up a two stream architecture based on AVOD for simultaneous 3D object detection and tracking; *(ii)* we introduce correlation features to capture object co-occurrences across time and perform frame-level detections in a high speed through interpolation; *(iii)* we utilize tracking result to improve detection performance and preliminary explore the algorithm for key frame selection.

2 Related Work

2.1 Video object detection

2.2 3D object detection

2.3 3D object tracking

3 Methodology

4 Experiments

5 Conclusion

References

[1] Aseem Behl, Despoina Paschalidou, Simon Donné, and Andreas Geiger. Pointflownet: Learning representations for 3d scene flow estimation from point clouds. *arXiv preprint arXiv:1806.02170*, 2018.

[2] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[3] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3038–3046, 2017.

[4] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.

[8] Philip Lenz, Andreas Geiger, and Raquel Urtasun. Followme: Efficient online min-cost flow tracking with bounded memory and computation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4364–4372, 2015.

[9] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Learning scene flow in 3d point clouds. *arXiv preprint arXiv:1806.01411*, 2018.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 408–417, 2017.

[13] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7210–7218, 2018.