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 [11, 12, 21] and their region based descendants [6, 10, 19], 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 [24, 25] 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

© 2019. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

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, 16] 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 [8], we transform AVOD [13] 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* [14].

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

Video object detection in image has received increased attention since ImageNet VID datasets introduced. Most approaches for video object detection utilize optical flows, which present temporal information in videos. Some representative work such as FGFA [24], it leverages temporal coherence on feature level, and improves the pre-frame features by aggregation of nearby features along the motion paths with the help of optical flows. Later a more efficient approach based on [24] has been presented in [25], it introduces three new techniques: sparsely recursive feature aggregation, spatially-adaptive partial feature updating and temporally-adaptive key frame scheduling, which make this unified approach faster, more accurate and more flexible.

There are also some approach try to learn temporal information between consecutive frames. D&T [8] set up a ConvNet architecture for simultaneous detection and tracking in video. In order to capture cross-occurrences across time, it aid a correlation operation in networks. Our Bi-AVOD architecture mainly inspired by this work.

2.2 3D object detection

Currently, most approaches in 3D object detection can be divided into three types: image based detectors, LiDAR based detectors and fusion based detectors.Image based approaches such as Mono3D[3], 3DOP[5] use camera data only, since image has limited depth information, specific hand-crafted geometric features are required. LiDAR based methods are usually done in two fronts, one is utilizing a voxel grid representation to encoder point cloud and applying 3D CNN for features extracted, these approaches including 3D FCN [15], Vote3Deep [7] and VoxelNet [23] *et al.*, these approaches suffer from the sparsity of point cloud and enormous computation cost in 3D convolution; others LIDAR based methods try to project point cloud to bird eye view (BEV) and apply 2D CNN for object detection, such as PIXOR[22], FaF[17] and Comple-YOLO [20] *et al.* These methods take advantage of the fact that objects in autonomous driving almost on the same plane thus loss of height information has little affect to the result, while the depth and Geometric information can be retained and computational complexity reduced significantly, making real-time detection possible. However, due to the sparsity of point cloud, the feature information after projecting is insufficient for accurate object detection especially for the small target.

There are also many multi-modal fusion methods that combine images and LiDAR data to improve detection accuracy. F-PointNet [18] first extracts the 3D bounding frustum of an object by extruding 2D bounding boxes from image detectors, then consecutively perform 3D object instance segmentation and amodal extent regression to estimate the amodal 3D bounding box. MV3D[4] extends the image based RPN of Faster R-CNN[19] to 3D and proposes a 3D RPN targeted at autonomous driving scenarios. MV3D uses every pixel in BEV feature map to multiple 3D anchors and then feeds the anchor to RPN to generate 3D proposals that are used to create view-specific feature crops from BEV feature maps and images. A deep fusion scheme is used to combine information from these feature crops to produce final detection output. However, MV3D does not work well for small targets due to the insufficient data for feature extracting caused by downsampling in convolutional feature extractors. AVOD[13] architecture is similar to MV3D in 3D RPN and feature fusion, however, its feature extract provides full resolution feature maps thus show greatly help in localization accuracy for small targets during the second stage of the detection framework. Our proposed architecture mostly based on AVOD mention above.

2.3 3D object tracking

More and more work has been done in 3D object tracking based on tracking by detection due to the rapidly development in 3D object detection. These approaches usually trend to apply 3D object detection and tracking simultaneously. FaF [17] jointly reasons about 3D detection, tracking and motion forecasting taking a 4D tensor created from multiple consecutive temporal frames. The most similar approach to our work is [9], however, their 3D detector is based on MV3D while ours is AVOD, and their detections association is done by solving a linear program after passing to a matching net and scoring net, while ours use a extending IOU based algorithm [2] by leveraging corresponding displacements over time.

3 Methodology

In this section we first present the problem statement of point cloud flow detection and tracking (Sec. 3.1), then we will give an overview of the Bi-AVOD approach (Sec. 3.2) that generates detections and tracklets give two adjacent key frames as input. We then introduce the correlation module (Sec. 3.3) that aid the network in the tracking process. Sec. 3.4 shows how we link detections to trajectories and Sec. 3.5 simply introduces evaluation metrics in multiple object tracking.

3.1 Problem formulation

Give a streaming point cloud of N frames $\{I_f\}$ for $f \in \{1,...N\}$, the stream object detection task requires a set of detections D_f for each frame I_f. Each detection set consists of object detections $\{D_f^i\}$ while $i \in \{1,...N_f\}$ (N_f is the number of detections in frame f).

3.2 Bi-AVOD model structure

3.3 Correlation module for object tracking

3.4 Link detections to trajectories

3.5 Metrics

4 Experiments

4.1 KITTI tracking datasets

4.2 Data preprocessing

4.3 Training and testing

4.4 Results

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] Erik Bochinski, Tobias Senst, and Thomas Sikora. Extending iou based multi-object tracking by visual information. *AVSS. IEEE*, 2018.

[3] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun. Monocular 3d object detection for autonomous driving. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2147–2156, June 2016. doi: 10.1109/CVPR.2016.236.

[4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[5] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals using stereo imagery for accurate object class detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1259–1272, 2018.

[6] 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.

[7] Martin Engelcke, Dushyant Rao, Dominic Zeng Wang, Chi Hay Tong, and Ingmar Posner. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1355–1361. IEEE, 2017.

[8] 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.

[9] Davi Frossard and Raquel Urtasun. End-to-end learning of multi-sensor 3d tracking by detection. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 635–642. IEEE, 2018.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Bo Li. 3d fully convolutional network for vehicle detection in point cloud. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1513–1518. IEEE, 2017.

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

[17] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018.

[18] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018.

[19] 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.

[20] Martin Simon, Stefan Milz, Karl Amende, and Horst-Michael Gross. Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds. In *European Conference on Computer Vision*, pages 197–209. Springer, 2018.

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

[22] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018.

[23] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.

[24] 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.

[25] 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.