3D Object Detection with Temporal Information

Tsung-Lin, Tsou
National Taiwan University
r10922081@csie.ntu.edu.tw
Yu-Jia, Liou
National Taiwan University
r10922083@csie.ntu.edu.tw
Yi-Syuan, Liou
National Taiwan University
r10944016@csie.ntu.edu.tw

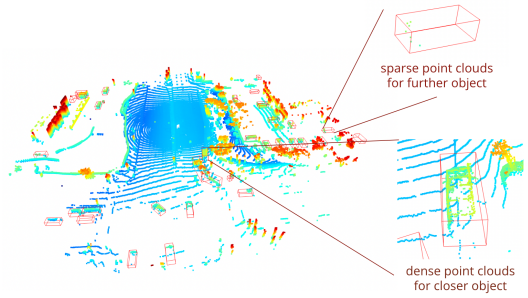

Figure 1. The visualization of different distance between objects and sensor.

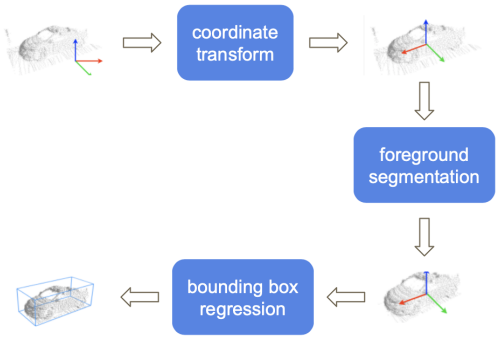

Figure 2. The model architecture of static tracks refinement.

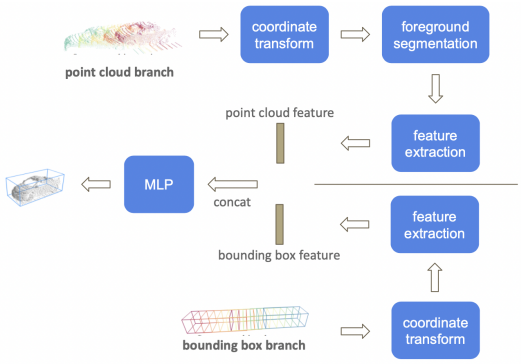

Figure 3. The model architecture of dynamic tracks refinement.

I. Introduction

With the advancement of 3D deep learning, there have been many studies on 3D object detection in recent years [SGJ+20][YZK21]. However, from Table I, we can observe most 3D object detectors have the problem that the longer the distance between object and sensor, the poorer the outcome turns on. In addition, the objects closer to the sensor have the dense point clouds, while the objects further from the sensor have the sparse point clouds (see Figure 1). This makes it more difficult for 3D object detector to identify the distant objects.

Therefore, our work is to follow the CVPR 2021 paper [QZN+21] and improve existed 3D object detector by obtaining temporal information. Specifically, since most of the work focuses on single-frame input, we can fuse multi-frame information and aggregate point clouds to get more complete data and improve performance.

II. Methodology

Our method is mainly based on 3DAL [QZN+21]. Since the code of 3DAL isn't available, our main contribution in this project is to use comparable components to reproduce it and prove the validity of temporal information via visualization. First, 3DAL uses 3D object detection module to obtain the initial 3D bounding boxes. Then, 3D object tracking module is applied to obtain temporal information for the purpose of densifying the LiDAR point clouds. Finally, it uses the more complete data, the densified LiDAR point clouds, to refine the 3D bounding boxes. We will elaborate each components of 3DAL in the following subsections: 3D object detection, 3D object

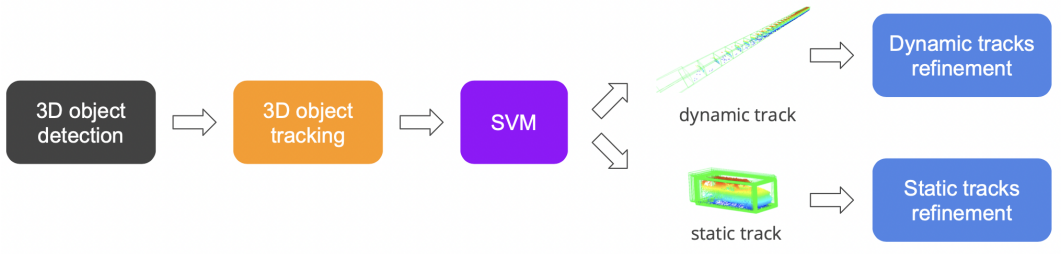

Figure 4. The overall pipeline of 3DAL.

tracking, motion state classification, static tracks refinement, and dynamic tracks refinement. The overall pipeline is shown in Figure 4.

A. 3D Object Detection and 3D Object Tracking

In this section, we simply use the CenterPoint [YZK21] to obtain static tracks and dynamic tracks as shown in Figure 4. First, CenterPoint voxelizes the LiDAR point clouds into small voxels across 3D space. Then, voxel feature extraction and 3D feature extraction is applied to obtain 2D BEV pseudo image feature map. Finally, in the first stage, it predicts the centers of 3D bounding boxes and regresses 3D information, which includes the velocity being used in 3D object tracking. In the second stage, it predicts the confidence scores and refines the estimates.

B. Motion State Classification

In this section, we use some heuristic features, i.e. begin-to-end distance, variance of center, and a support vector machine (SVM) to classify the object tracks obtained by 3D object detection and 3D object tracking. Albeit the model is simple, the accuracy of classification can be as high as 97 percent. (Note: All components in this section are implemented by ourself.)

C. Static Tracks Refinement

For the static tracks classified by SVM, we use static tracks refinement model, as shown in Figure 2, to regress the final 3D bounding boxes in static tracks. More specifically, we only need to predict a single box for each static track and it can be transformed to each frame within the static track through the known sensor poses. First, we transform the object points to a box coordinate before the per-object processing as in 3DAL [QZN+21], such that the point clouds are more aligned across objects. In the box coordinate, the +X axis is the box heading direction, the origin is the box center. Then, the object points are passed through an segmentation network to segment the foreground. Finally, the foreground object points will be regressed by a PointNet based bounding box regression network. (Note: All components in this section are implemented by ourself.)

	Vehicle	Pedestrian	Cyclist	mAPH
$(0, 30)$	0.902	0.773	0.799	0.825
$[30, 50)$	0.698	0.696	0.671	0.688
$[50, +\infty)$	0.441	0.574	0.535	0.517

Table I
PERFORMANCE OF CENTERPOINT[YZK21] ON WOD VAL.

D. Dynamic Tracks Refinement

For the dynamic tracks classified by SVM, we use dynamic tracks refinement model, as shown in Figure 3, to regress the final 3D bounding boxes in dynamic tracks. More specifically, we leverage both the point cloud sequence and the bounding box sequence without aligning object points to a keyframe explicitly as in 3DAL [QZN+21] to predict different 3D bounding boxes for each frame within the dynamic track. For the point cloud branch, we first transform the object points to a box coordinate and segment the foreground similar to the static one. Then we use a PointNet based model to extract point cloud feature. For the bounding box branch, we use similar method to extract bounding box feature. Finally, the features are concatenated to form the joint feature which will be passed through a PointNet based box regression network to predict the final 3D bounding boxes. (Note: All components in this section are implemented by ourself.)

III. EXPERIMENT

A. Results

As shown in Table II and III, our method outperforms CenterPoint in all terms. Furthermore, we also get better accuracy than 3DAL by about 0.01 and 0.001 on static and dynamic tracks, respectively. However, in Table IV, due to the use of test-time augmentation in [QZN+21], their work gets higher performance than ours.

In addition, we find that dynamic object has higher 3D accuracy in comparison with static object. The main reason may be that the dynamic objects are often closer to the sensor and thus have denser point cloud.

	2D IoU	3D IoU	3D acc
CenterPoint	0.858	0.768	0.774
Ours	0.869	0.785	0.835
3DAL	-	-	0.823

Table II
STATIC TRACKS REFINEMENT.

	2D IoU	3D IoU	3D acc
CenterPoint	0.764	0.703	0.774
Ours	0.780	0.717	0.858
3DAL	-	-	0.857

Table III
DYNAMIC TRACKS REFINEMENT.

B. Visualization

In Figure 5 and Figure 6, we could see that when the lidar points are sparse, the model that utilizes temporal information has more accurate detection.

IV. SUMMARY

We show that consider point clouds in multiple frames do help improve the 3D object detection, especially in far objects with sparse point cloud in single frame. Another interesting point is that we find that dynamic object has higher 3D accuracy in comparison with static object. By replacing some components in 3DAL [QZN+21], we show that our performance are better than the original paper in accuracy in dynamic tracks, which means that leverage the information of adjacent frames do help detection and are not specific to certain model structure.

The further improvement of this pipeline could be modifying the input of box regression network in dynamic tracks. Instead of feeding the whole track into the network, we could try to estimate the velocity of the object and map all the point cloud in other frames to the first one. By solving the alignment error, we may get a better 3D object detection in dynamic tracks.

A. Division of Work

Tsung-Lin, Tsou: Survey, building modules with temporal information, report.
Yu-Jia, Liou: Survey, train 3D detection model, analysis with visualization and report.
Yi-Syuan, Liou: Survey, train 3D detection model, analysis with visualization and report.

	Detection	Temporal	TTA	Veh	Ped
Ours	✓			0.767	0.790
Ours	✓	✓		0.789	0.812
3DAL	✓			0.746	0.780
3DAL	✓	✓	✓	0.845	0.829

Table IV
AP ON WAYMO OPEN DATASET VAL SET.

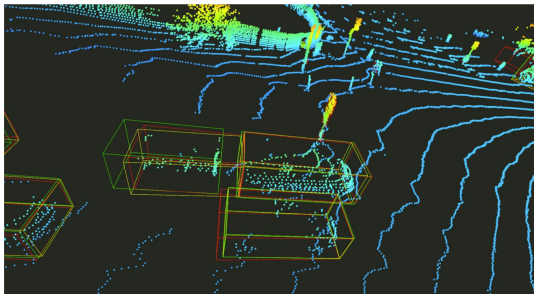

Figure 5. Visualization of different models in detection. RED: ground truth, YELLOW:w temporal, GREEN:wo temporal

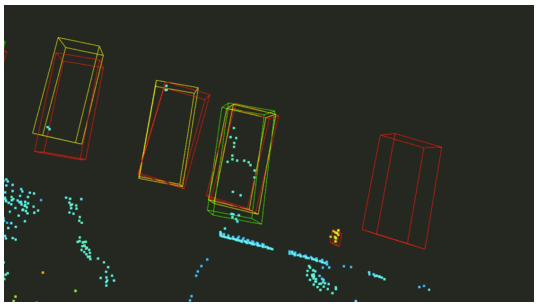

Figure 6. Visualization of different models in detection. RED: ground truth, YELLOW:w temporal, GREEN:wo temporal

REFERENCES

[QZN+21] Charles R Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3d object detection from point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6134–6144, 2021.

[SGJ+20] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020.

[YZK21] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.