Technical Report for CVPR 2023 Workshop on Argoverse2 3D Object Tracking Competition

Jae-Keun Lee[1,3], Jin-Hee Lee[2], Joohyun Lee[2], Heechul Jung[3], Soon Kwon[1,2]
[1]FutureDrive, [2]DGIST (Daegu Gyeongbuk Institute of Science and Technology),
[3]KNU (Kyungpook National University)
{lejk8104,soon.kwon}@futuredrive.net, {jhlee07,jhlee0714}@dgist.ac.kr, heechul@knu.ac.kr

Abstract

LiDAR-based 3D Multi-Object Tracking (MOT) technology is widely researched in various fields such as autonomous driving and robotics. However, the development of deep learning-based MOT techniques requires a large-scale labeled dataset. This is because training deep learning models heavily relies on abundant labeled data that covers diverse objects in real-world road environments. Nevertheless, the process of collecting and annotating this data is time-consuming and costly. Fortunately, the recently released Argoverse2 dataset offers point cloud and 3D annotation data for 26 different classes. Therefore, researchers can utilize this dataset to develop and evaluate models in the field of Lidar-based 3D MOT. In this report, we propose a solution for 3D MOT by improving the detection and tracking modules of CenterPoint-VoxelNet. The proposed model achieves high accuracy, surpassing the baseline models, and ranks second in the 2023 Argoverse 3D Object Tracking Competition Leaderboard.

1. Methods

1.1. Overviews

Our model sequentially processes multiple LiDAR point cloud frames as input and predicts object categories and states, bounding box location and size, and tracker ID. As shown in Figure 1, the proposed model consists of the 3D Detector and 3D Tracker modules, each of which is described in detail in Section 1.2 and Section 1.3.

1.2. 3D Detector

The proposed 3D Object Detector module first receives point cloud data of each frame as input data and computes various information for each object. This information includes the object's class, classification score, predicted bounding box (center x, center y, center z, width, length, height, yaw), and predicted IoU. Then, this information is combined to detect the objects within each frame. The following outlines the processing sequence of the 3D Object Detector module.

Multi-Frame Point Cloud: This model takes as input the merged point cloud between the current frame and the previous frame. During the merging process, the point cloud from the previous frame is transformed into the current frame's coordinate system by applying the ego vehicle's pose information.

Voxel Feature Encoding: To perform efficient object detection on large-scale point clouds, we employ Voxel Feature Encoding (VFE) as a next step. VFE groups the points within predefined voxel regions and applies a Multi-Layer Perceptron (MLP) to generate a sparse voxel tensor, which sends as the input data for the backbone.

Backbone and neck: The model utilizes an improved CenterPoint-VoxelNet [6] backbone and neck network to extract rich semantic and spatial features. These features are transferred to the Multi-Group Head for 3D object detection.

Multi-Group Head: Following the baseline model, we adopted the strategy of a Multi-Group Head. This head consists of six sub-heads, which perform feature-wise classification and feature-wise regression to predict the attributes and locations of objects. Specifically, the first sub-head targets Regular vehicle, while the second sub-head includes objects like Pedestrian, Bicyclist, Motorcyclist, and Wheeled rider. The third sub-head predicts static objects such as Bollard, Construction cone, Sign, Construction barrel, Stop sign, and Mobile pedestrian crossing sign. The fourth sub-head focuses on large vehicle, including Bus, Box truck, and Truck. The fifth sub-head includes small objects such as Bicycle, Motorcycle, Wheeled device, Wheelchair, and Stroller. Finally, the sixth sub-head concentrate on the Dog class.

1.3. 3D MOT

The proposed 3D Tracker module is designed based on AB3DMOT [4]. This module leverages sequential object

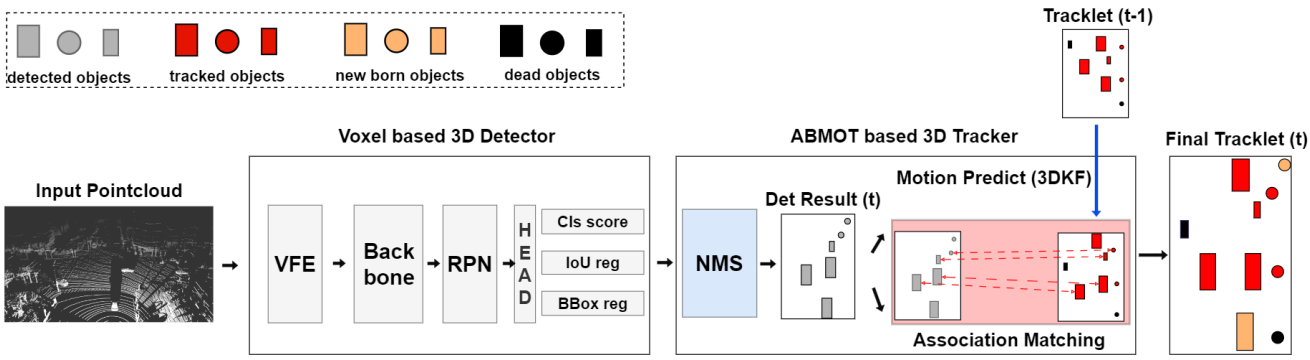

Figure 1. **3D MOT pipeline.** Our model is divided into detection and tracking modules, which perform tracking based on information from objects detected in previous and current frame. The detected objects in current frame are described as gray color and the tracked objects are presented as red color. In addition, the newly added objects are shown as orange color and dead objects with a low confidence score are represented as black color.

detection results as input data. By employing techniques such as Non-Maximum Suppression (NMS), 3D Kalman Filter (3DKF) based motion prediction, and Data Association matching, it predicts the object's tracker IDs, states, and bounding box information.

Non-Maximum Suppression: We enhance the NMS to supply the Tracker module with higher quality bounding boxes. i.e, we redefine the confidence score by simultaneously considering both IoU and classification scores. After sorting these confidence scores, we use circle NMS to eliminate duplicated bounding boxes.

Motion Predict: We apply a 3D Kalman Filter (3DKF) to the bounding boxes suppressed through the NMS post-processing step. The 3DKF estimates the trajectories of the tracklets in the current frame, including as input data the location, bounding box size, and yaw information of tracklets from the previous frame. Following baseline model LT3D [2], we convert the bounding box coordinates of each tracklet from lidar frame coordinates to world frame coordinates in the 3DKF.

Association Matching: To generate the association matrix, we estimate IoU between the bounding boxes predicted by 3DKF and the detection results from the current frame. Subsequently, we employ Hungarian matching algorithm to minimize costs. During this process, tracklets with an IoU highter than a predefined threshold (e.g., 0.1) survive, while unmatched detection results are defined as new tracklets. Conversely, tracklets with an IoU lower than the threshold are filtered out.

2. Dataset and Metrics

2.1. Dataset

In this competition, we use the Argoverse 2 Sensor Dataset [5]. This dataset consists of 700 training sequences, 150 validation sequences, and 150 test sequences. Each sequence includes synchronized LiDAR and camera data, 3D annotation data, and map information. We primarily focused on the LiDAR point cloud data and 3D annotation data for this challenge.

2.2. Metric

In this competition, performance is assessed by considering objects from 26 different categories, including vehicles, pedestrians, and motorcycles, within a range of 50 meters. The official evaluation metrics for this challenge are Higher Order Tracking Accuracy (HOTA) [1], Average Multi-Object Tracking Accuracy (AMOTA) [3], and Multiple Object Tracking Accuracy (MOTA). For a more detailed description of these metrics, refer to the official competition website.

3. Experiment

3.1. Implementation detail

Our model was trained for 6 epochs using 4 NVIDIA RTX A6000s with a batch size of 4. We adopted an initial learning rate of 0.003, utilized the ADAM optimizer to update learnable parameters, and set the weight decay to 0.01. The voxel size was set to [0.1m, 0.1m, 0.15m], and the detection range covered [-60m, -60m, -3m, 60m, 60m, 3m].

3.2. Results

Table 1 shows the evaluation results of our model on the validation dataset of Argoverse2. Our model achieves 44.29 HOTA, 17.07 AMOTA, and 33.87 MOTA on this validation dataset. Similarly, as shown in Table 2, the model achieves 44.36 HOTA, 17.47 AMOTA, and 32.61 MOTA on the test dataset. These results outperform the performance of the baseline model by 4.38 HOTA. As a result, our model ranked second on the 2023 Argoverse Tracking Competition Leaderboard.

Method	HOTA	AMOTA	MOTA
AIDrive (ours)	**44.30**	**17.08**	**32.87**

Table 1. Evaluation results on the Argoverse2 validation dataset.

Method	HOTA	AMOTA	MOTA
Le3DE2E	56.19	19.53	39.34
AIDrive (ours)	**44.36**	**17.47**	**32.61**
dgist-cvlab	41.49	7.88	17.97
Baseline	39.98	7.1	16.21

Table 2. Results on test dataset for the 2023 Argoverse2 Tracking Competition Leaderboard.

References

[1] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, 129:548–578, 2021. 2

[2] Neehar Peri, Achal Dave, Deva Ramanan, and Shu Kong. Towards long-tailed 3d detection. In *Conference on Robot Learning*, pages 1904–1915. PMLR, 2023. 2

[3] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10359–10366, 2020. 2

[4] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics. In *Eur. Conf. Comput. Vis. Worksh.*, 2020. 1

[5] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 2

[6] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 1