Le3DE2E Solution for AV2 2024 Unified Detection, Tracking, and Forecasting Challenge

Feng Chen
Lenovo Research
chenfeng13@lenovo.com

Kanokphan Lertniphonphan
Lenovo Research
klertniphonp@lenovo.com

Yaqing Meng
Lenovo Research
mengyq3@lenovo.com

Ling Ding
Lenovo Research
dingling3@lenovo.comn

Jun Xie
Lenovo Research
xiejun@lenovo.com

Kaer Huang
Lenovo Research
huangke1@lenovo.com

Zhepeng Wang
Lenovo Research
wangzpb@lenovo.com

Abstract

This report presents our team's 'Le3DE2E' solution for the AV2 2024 Unified Detection, Tracking, and Forecasting Challenge at Workshop on Autonomous Driving (WAD), CVPR2024. The main goal of the challenge is to precisely detect, track, and forecast 26 object categories in end-to-end perception. Since object detection plays a crucial role in the end-to-end system, our primary focus has been on enhancing object detection performance. We introduce an object detection network that includes a linear kernel backbone [2], a heatmap encoder, and a deformable decoder [1]. We achieved 1st place in detection and tracking challenges and 2nd in forecasting challenges at the CVPR 2024 WAD.

1. Introduction

The task assesses end-to-end perception tasks on detection, tracking, and multi-agent forecasting using the Argoverse 2 sensor dataset [9]. The dataset includes track annotations for 26 object categories. During testing, our algorithm detects objects in the present frame, tracks object trajectories, and predicts trajectories for the subsequent 3 seconds. This holistic task differs from motion forecasting as it lacks provided tracking ground truths.

2. Method

The system overview is illustrated in figure 1. Object detection plays a vital role in our end-to-end system, emphasizing improved detection performance by enhancing feature extraction and the detection head while following the baseline [4] for tracking and forecasting.

2.1. Detection

Our detection system consists of two primary components. In the backbone, we implement the LinK [2] method for more extensive spatial feature extraction using convolution. Initially, weights are assigned to non-empty regions through a linear kernel generator. Subsequently, the pre-computed aggregation results from the overlapped blocks are reused.

In the detection head, we employ FocalFormer3D [1] to reduce false negatives in object detection. The multi-stage heatmap encoder utilizes Hard Instance Probing (HIP). Positive instances are suppressed to focus on false negatives at each stage to enhance overall recall. Box-level queries are sent to Deformable DETR [11] and the object queries are forwarded to the MLP classifier.

2.2. Tracking and Forecasting

AB3DMOT tracker [8] is utilized to process object detection outcomes. This approach combines a 3D Kalman filter and a Hungarian algorithm to match objects across frames. Subsequently, LSTM is employed for predicting trajectories within the next 3 seconds.

2.3. Test Time Augmentation and Ensemble

During the inference stage, Test Time Augmentation (TTA) is implemented to enhance performance further. Moreover, Non-Maximum Suppression (NMS) is used to consolidate the results obtained from augmented inputs.

Weighted Box Fusion (WBF) [6] is employed to combine multiple results from models with varying training configurations to enhance detection accuracy. The detection bounding boxes are clustered based on intersection-over-union (IoU), and subsequently, fused box coordinates were calculated as the weighted average of the merged boxes.

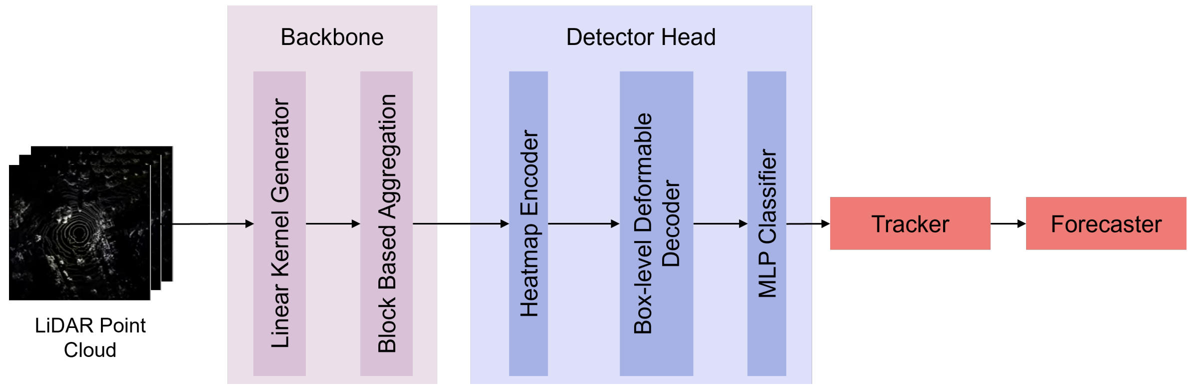

Figure 1. The system overview.

3. Experiment

3.1. Dataset and Evaluation Metric

The competition utilized the Argoverse 2 Sensor Dataset [9], comprising 1000 scenes, totaling 4.2 hours of driving data. Each vehicle log spans roughly 15 seconds and contains an average of 150 LiDAR scans captured at 10 FPS. Additionally, the dataset features 7 surrounding cameras recording at 20 FPS. For the E2E Forecasting track, one keyframe is sampled at 2Hz from the training, validation, and testing sets.

Detection Composite Detection Score (CDS) is used in the challenge, which evaluates precision, recall, object extent, translation error, and orientation concurrently. The mean metrics are derived as an average across 26 distinct object categories.

Tracking HOTA[3] is the key metric for the challenge, while AMOTA and MOTA are significant secondary metrics for reference. HOTA offers a balanced assessment of accurate detection, association, and localization within a single unified metric. MOTA incorporates false positives, missed targets, and switches to calculate tracking accuracy, while AMOTA considers the confidence of predicted tracks by averaging over all recall thresholds.

Forecasting The primary evaluation metric includes Forecasting mAP (mAP_F)[5], ADE, and FDE, which are averaged across both static, and non-linearly moving cohorts. mAP_F is the key metric for the challenge, which defines a true positive when a positive match occurs in both the current timestamp (T) and the future (T+N). ADE represents the average L2 distance between the best-forecasted trajectory and the ground truth, whereas FDE measures the L2 distance between the endpoint of the best-forecasted trajectory and the ground truth.

The evaluation of Detection is within 150 meters range while Tracking and Forecasting are within 50 meters range.

	mCDS(↑)	mAP(↑)
Tranfusion [10] (baseline)	0.42	0.50
FocalFormer3D [1]	0.48	0.58
FocalFormer3D + LinK [2]	0.49	0.58
FocalFormer3D + LinK + TTA	0.52	0.61

Table 1. An Ablation study on object detection

3.2. Implementation Details

We first voxelize the point clouds and utilize LinK for voxel encoding. Subsequently, we employ SECOND as the backbone and a convolution layer as the neck to transform the voxel feature into a Bird's Eye View (BEV) feature. The voxel size for the LiDAR encoder is (0.075m, 0.075m, 0.2m) across all tasks. Specifically, the point cloud range is restricted to [-54m, 54m] x [-54m, 54m] x [-3m, 3m] to cover the maximum range in tracking and forecasting. For the detection, the point clouds are constrained within [-153.6, -153.6, -5.0, 153.6, 153.6, 3.0]. In the LiDAR backbone, we down-sample voxels to 1/8.

Training We trained the detector for 20 epochs using the AdamW optimizer, with a learning rate of 1e-4, weight decay of 0.01, and a total batch size of 16 on 8 x V100 GPUs. Employing cyclic annealing to decay the learning rate, Class-Balanced Grouping and Sampling (CBGS) was used in the first 15 epochs and then disabled in the last 5 epochs. The ablation test results on validation can be found in table 1.

TTA and Ensemble Each model underwent global scaling with [0.95, 1, 1.05] and flipping for the xz-plane and yz-plane for TTA. Multiple models were trained with three voxel sizes of [0.05m, 0.075m, 0.1m], with or without CBGS augmentation. We combined the results with our previous year's end-to-end model [7] to generate the final results.

Team	mCDS(↑)	mAP(↑)	mATE(↓)	mASE(↓)	mAOE(↓)
Le3DE2E (Ours)	**0.43**	**0.52**	**0.36**	**0.27**	**0.38**
BEV	0.37	0.46	0.40	0.30	0.50
Detectors	0.34	0.42	0.39	0.30	0.50
Valeo3Cast	0.31	0.4	0.41	0.3	0.8
Anony_3D	0.31	0.39	0.43	0.32	0.6
Baseline	0.14	0.18	0.49	0.34	0.72

Table 2. 3D Object Detection Leaderboard

Team	HOTA(↑)	AMOTA(↑)	MOTA(↑)
Le3DE2E (Ours)	**64.60**	**26.32**	**51.27**
Valeo4Cast	61.39	24.06	47.83
Anony_3D	44.36	17.47	32.61
dgist_cvlab	41.49	7.88	17.97
Baseline	39.98	7.1	16.21

Table 3. Tracking Leaderboard on End-to-End Forecasting Challenge

Team	mAP_F(↑)	ADE(↓)	FDE(↓)
Valeo4Cast	**63.82**	**2.14**	**2.43**
Le3DE2E (Ours)	50.53	4.07	4.60
dgist-cvlab	45.83	4.09	4.53
Baseline	14.51	5.1	7.32

Table 4. Forecasting Leaderboard on End-to-End Forecasting Challenge

4. Conclusion

In this challenge, we improved the object detection module by integrating the LinK backbone and FocalFormer 3D, resulting in enhanced detection results. Our solution was evaluated across three sub-challenges: Detection, Tracking, and Forecasting. In the 3D Object Detection category, table 2 shows our solution achieving 0.43 mCDS, ranking 1st place in Detection. Table 3 presents the final Tracking leaderboard, with our solution obtaining 64.60 HOTA, ranking 1st. In the Forecasting task, as shown in table 4, our solution achieved 50.53 mAP_F, ranking the 2nd place.

References

[1] Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, and Jose M Alvarez. Focalformer3d: Focusing on hard instance for 3d object detection. 2023. 1, 2

[2] Tao Lu, Xiang Ding, Haisong Liu, Gangshan Wu, and Limin Wang. Link: Linear kernel for lidar-based 3d perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1105–1115, 2023. 1, 2

[3] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip H. S. Torr, Andreas Geiger, Laura Leal-Taixé, and B. Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129:548 – 578, 2020. 2

[4] Deva Ramanan Shu Kong Neehar Peri, Achal Dave. Towards long-tailed 3d detection. In *Conference on Robot Learning*, 2022. 1

[5] Neehar Peri, Jonathon Luiten, Mengtian Li, Aljovsa Ovsep, Laura Leal-Taix'e, and Deva Ramanan. Forecasting from lidar via future object detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17181–17190, 2022. 2

[6] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, pages 1–6, 2021. 1

[7] Zhepeng Wang, Feng Chen, Kanokphan Lertniphonphan, Siwei Chen, Jinyao Bao, Pengfei Zheng, Jinbao Zhang, Kaer Huang, and Tao Zhang. Technical report for argoverse challenges on unified sensor-based detection, tracking, and forecasting. *ArXiv*, abs/2311.15615, 2023. 2

[8] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. AB3DMOT: A Baseline for 3D Multi-Object Tracking and New Evaluation Metrics. *ECCVW*, 2020. 1

[9] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021. 1, 2

[10] Xinge Zhu Qingqiu Huang Yilun Chen Hongbo Fu Xuyang Bai, Zeyu Hu and Chiew-Lan Tai. TransFusion:

Robust Lidar-Camera Fusion for 3d Object Detection with Transformers. *CVPR*, 2022. 2

[11] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1