Supplementary Materials of Time3D: End-to-End Joint Monocular 3D Object Detection and Tracking for Autonomous Driving

Peixuan Li
SAIC PP-CEM
lipeixuan@saicmotor.com

Jieyu Jin
SAIC PP-CEM
jinjieyu@saicmotor.com

1. Monocular 3D Object Detection.

We design our monocular 3D object detector following the "joint detection and tracking" paradigm, FairMOT [2]. As shown in Fig. 2, We adopt FairMOT encoder-decoder as the main structure to our 3D detector, which applies a DLA-34 as the backbone to fuse multi-layer features. We employ a KM3D head to detect a 3D box, an anchor-free 3D detector, to provide fair features for detection and Re-ID. Specifically, five parallel heads are appended to FairMOT encoder-decoder to estimate object 2D main center, keypoints, dimension, orientation, and 3D score. The geometry reasoning module with camera intrinsic is applied to compute a 3D position. In the Re-ID branch, we apply a convolution layer with 256 kernels on top of backbone features to extract re-ID features for each location in the main center. KM3D branch is explicitly supervised by ground truth, and the Re-ID branch is implicitly supervised by spatial-temporal information flow.

In the ablation learning of non-end-to-end learning, we need to separate the 3D detector and Re-ID extractor. We simply employ two FairMOT encoder-decoder networks for 3D detection and Re-ID embedding extracting, respectively. We train the 3D detector following the KM3D pipeline [1]. We train the Re-ID branch only through tracking head in spatial-temporal information flow with other cues.

2. More Qualitative Results .

More qualitative results are shown in Fig. 3,Fig. 5,Fig. 4, including day and night, and different weather conditions.

3. More Ablation Studies .

We design the Temporal-Consistency Loss to constrain the trajectory smoothness in the training phase. This allows us to smooth the trajectory without any additional calculations and post-processing. It can be seen in the qualitative analysis that the trajectories generated by Time3D are smoother. Fig. 1 shows 3D corner error and Temporal-Consistency error with and without Temporal-Consistency loss. It can be seen that Time3D with Temporal-Consistency loss reduces the temporal-Consistency error and does not affect the position error of independent objects in a single frame. The detection accuracy and temporal-consistency error in Tab. 1 also demonstrate this effect.

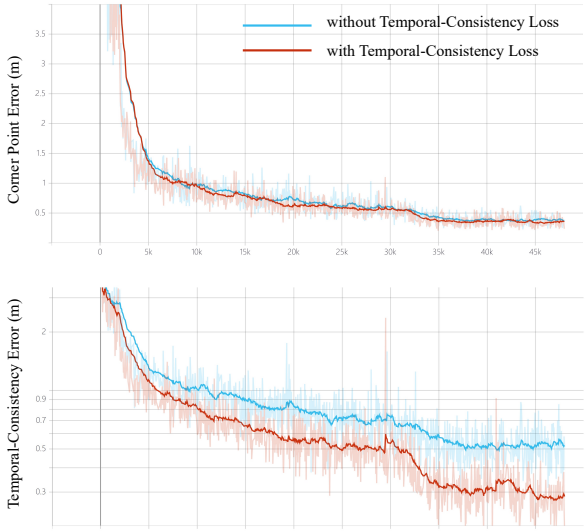

Figure 1. **Ablation Studies For Temporal-Consistency Loss.** Training on Nuscenes Training set. Corner Point Error denotes 3D corners distance between prediction and Ground Truth in a single frame. Temporal-Consistency Error denotes 3D corners distance between prediction and Ground Truth about the 3D corner distance of the same object in different frames.

Table 1. **Ablation for Temporal-Consistency Loss.** TC denotes Temporal-Consistency Error.

Setting	TC ↓ (m)	mAP (%) ↑	mATE (m) ↓	mASE (1-iou)↓	mAOE (rad)↓	NDS (%)↑
w/o	0.51	31.1	0.730	0.249	0.511	39.4
w/	0.28	31.2	0.732	0.254	0.504	39.4

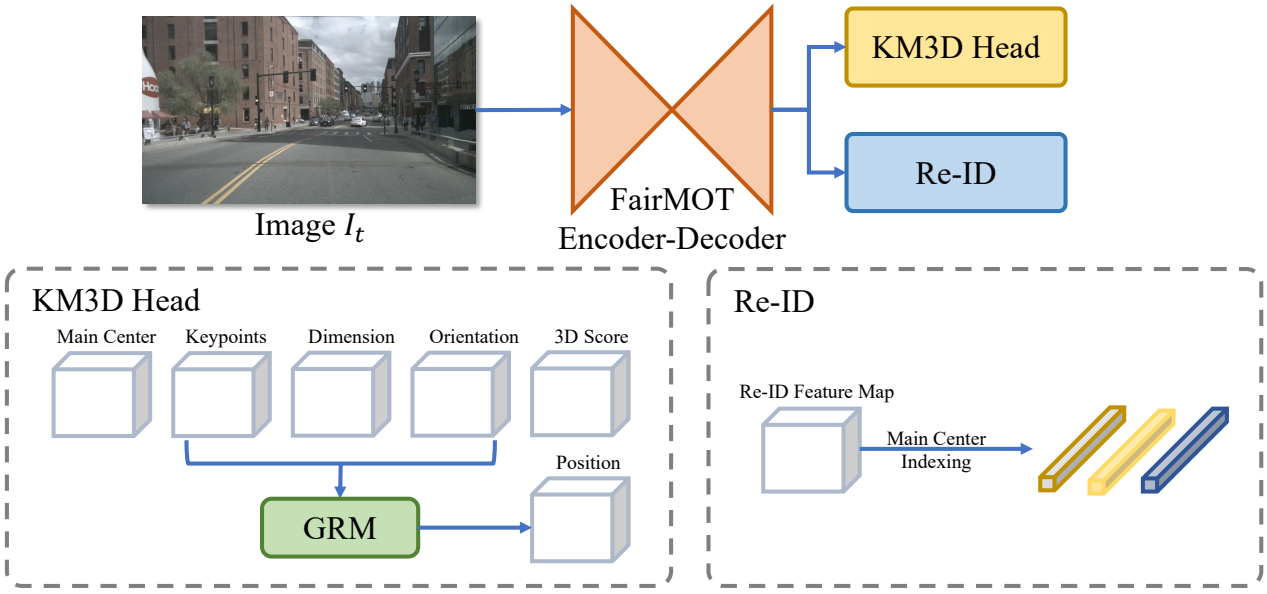

Figure 2. **Overview of Our Monocular 3D Detector.** An multi-scale FairMOT encoder-decoder network and anchor-free 3D detection head are used in our framework to provide fair features for both detection and Re-ID.

References

[1] Peixuan Li. Monocular 3d detection with geometric constraints embedding and semi-supervised training. 2020. 1

[2] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.*, 129(11):3069–3087, 2021. 1

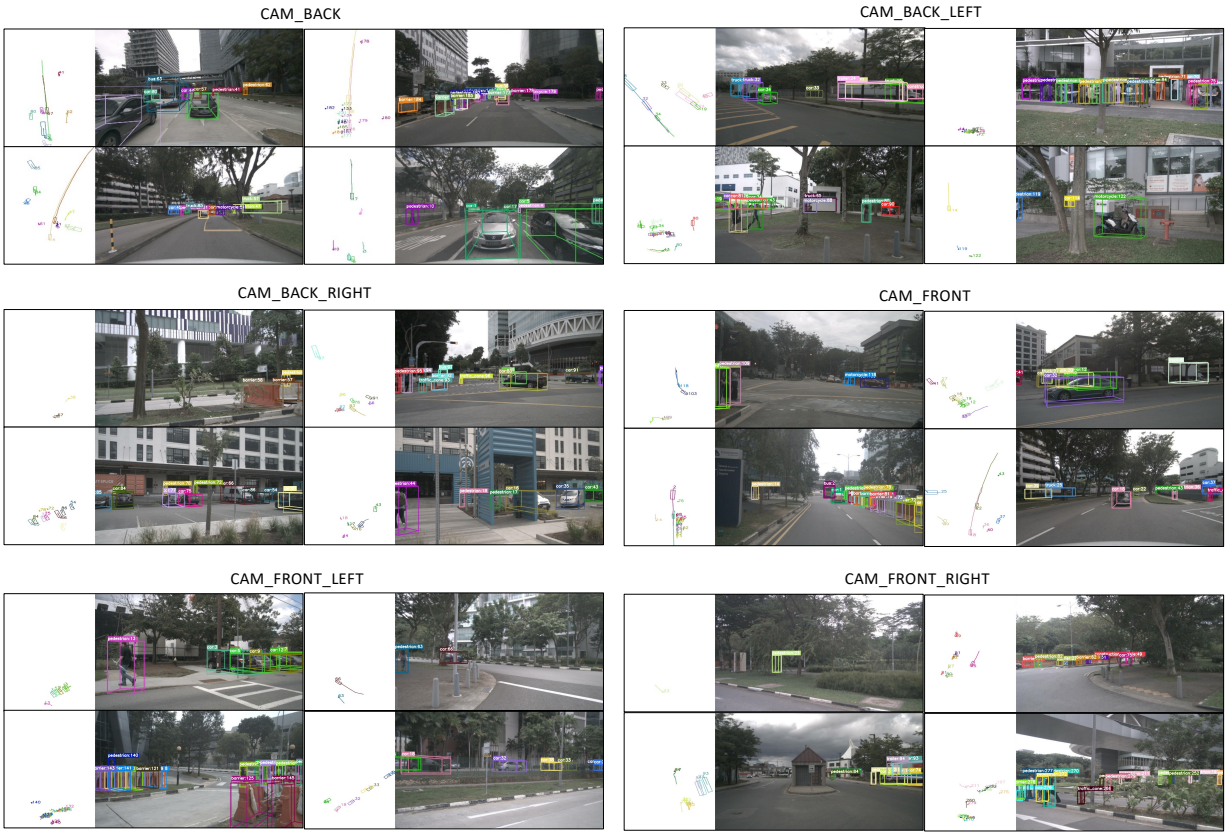

Figure 3. **Qualitative Results: Sunny Day**

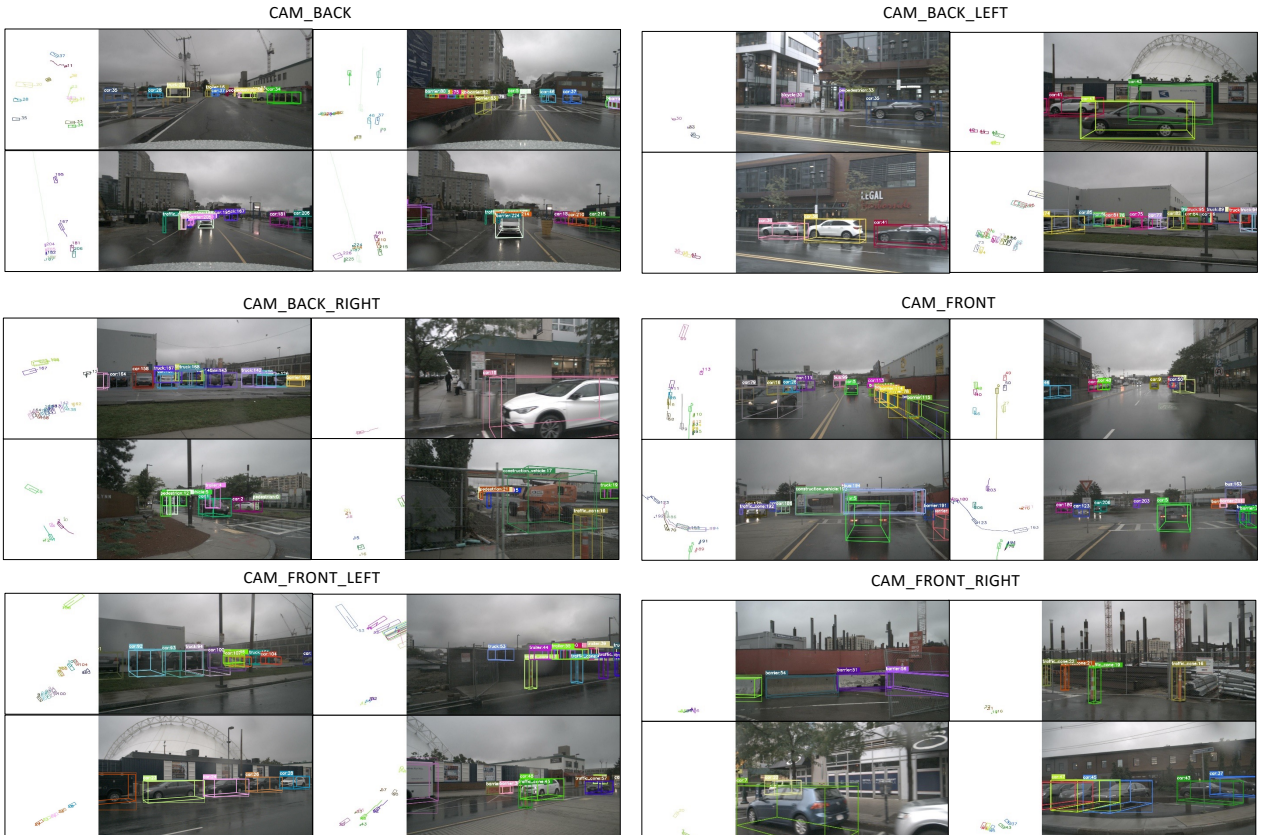

Figure 4. **Qualitative Results: Rainy Day**

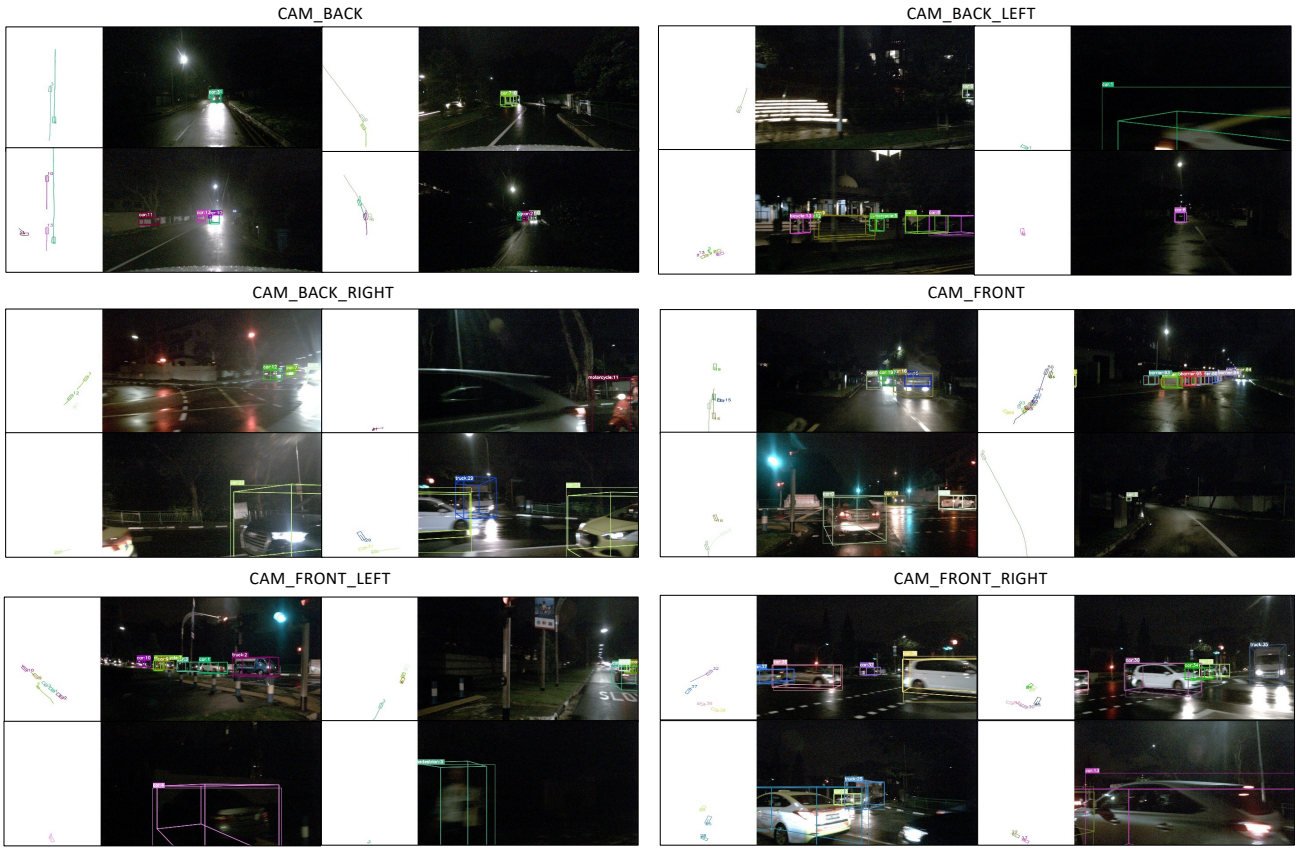

Figure 5. **Qualitative Results: Night**