Delving into Motion-Aware Matching for Monocular 3D Object Tracking Supplementary Material

Kuan-Chih Huang[1] Ming-Hsuan Yang[1,2,3] Yi-Hsuan Tsai[2]

[1]University of California, Merced [2]Google [3]Yonsei University

1. Main Evaluation Metrics

The nuScenes dataset evaluates the 3D MOT performance mainly by utilizing AMOTA, which is built upon the sAMOTA (scaled AMOTA) metric [10] to deal with the problem of MOTA [1] that may tend to filter low-confidence detections because of the potential of causing false-positive results. The AMOTA is defined as MOTA [1] over n recall thresholds:

$$\text{AMOTA} = \frac{1}{n} \sum_{r \in \{\frac{1}{n}, \frac{2}{n}, \ldots, 1\}} \text{MOTAR},$$
$$\text{MOTAR} = \max(0, 1 - \frac{\text{IDS}_r + \text{FP}_r + \text{FN}_r - (1-r) * \text{GT}}{r * \text{GT}}), \tag{1}$$

where IDS_r, FP_r, FN_r denote the number of identity switches, false positives, and false negatives calculated at the certain recall r, and GT is the number of ground truth.

2. More Experimental Results

Analysis of different frame lengths (T). In Table 1, we investigate the effect of different frame lengths utilized in our motion transformer. It is worth noting that we apply the global representation instead of the motion representation for the case of the single frame ($T = 1$). We find that, as the frame number becomes larger, the performance is improved gradually, especially when using more than 4 frames. In the main paper, we use $T = 6$ as our final setting.

Effectiveness of time positional encoding in motion transformer. In Table 2, we show that using a learnable time positional encoding for the proposed motion transformer improves the performance since it makes the model aware of motion cues at different timestamps.

3D MOT results on the nuScenes validation set. Table 3 and Table 4 present the 3D tracking performance on the nuScenes validation set for single-camera and multi-camera settings. It shows that our MoMA-M3T achieves better results than existing methods on both tracking settings, which validates the effectiveness of our approach.

	Frame Number	AMOTA↑	AMOTP↓	MOTA↑
(a)	1	29.5	1.447	25.5
(b)	2	30.2	1.441	26.2
(c)	4	30.9	1.436	**27.1**
(d)	6	**31.1**	**1.432**	**27.1**

Table 1. **Analysis of different frame lengths for our motion transformer** on the nuScenes validation set. We use the global representation to deal with the single frame observation.

Setting	AMOTA↑	AMOTP↓	MOTA↑
w/o Time Positional Enc.	30.7	1.435	26.4
w/ Time Positional Enc.	**31.1**	**1.432**	**27.1**

Table 2. **Effectiveness of time positional encoding** in motion transformer on the nuScenes validation set.

Method	AMOTA↑	AMOTP↓	RECALL↑	MOTA↑	MOTP↓
CenterTrack [13]	6.8	1.54	0.23	6.1	-
TraDeS [11]	11.8	1.48	0.23	-	-
PermaTrack [8]	10.9	-	-	8.1	-
DEFT [2]	20.9	-	-	17.8	-
Time3D [6]	26.0	**1.38**	-	20.7	0.82
QD-3DT [5]	24.2	1.518	0.399	21.8	0.81
MoMA-M3T (Ours)	**31.1**	1.432	**0.468**	**27.1**	**0.766**

Table 3. **3D MOT performance on the nuScenes validation set for the single-camera tracking setting.** We use **bold** to highlight the best results.

Method	AMOTA↑	AMOTP↓	RECALL↑	MOTA↑	MOTP↓
MUTR3D [12]†	29.4	1.498	0.427	26.7	0.799
CC-3DT [3]◇	42.9	1.257	0.538	35.7	-
MoMA-M3T (Ours)†	36.2	1.369	0.484	31.2	0.794
MoMA-M3T (Ours)◇	**44.8**	**1.225**	**0.550**	**38.8**	**0.714**

Table 4. **3D MOT performance on the nuScenes validation set for the multi-camera tracking setting.** † and ◇ denote using DETR3D [9] and BEVFormer [7] as the detector with the ResNet101 [4] backbone, respectively.

3. Qualitative Visualization

More visualization results. In Figure 1, we show example visualization results on the nuScence validation set. It can

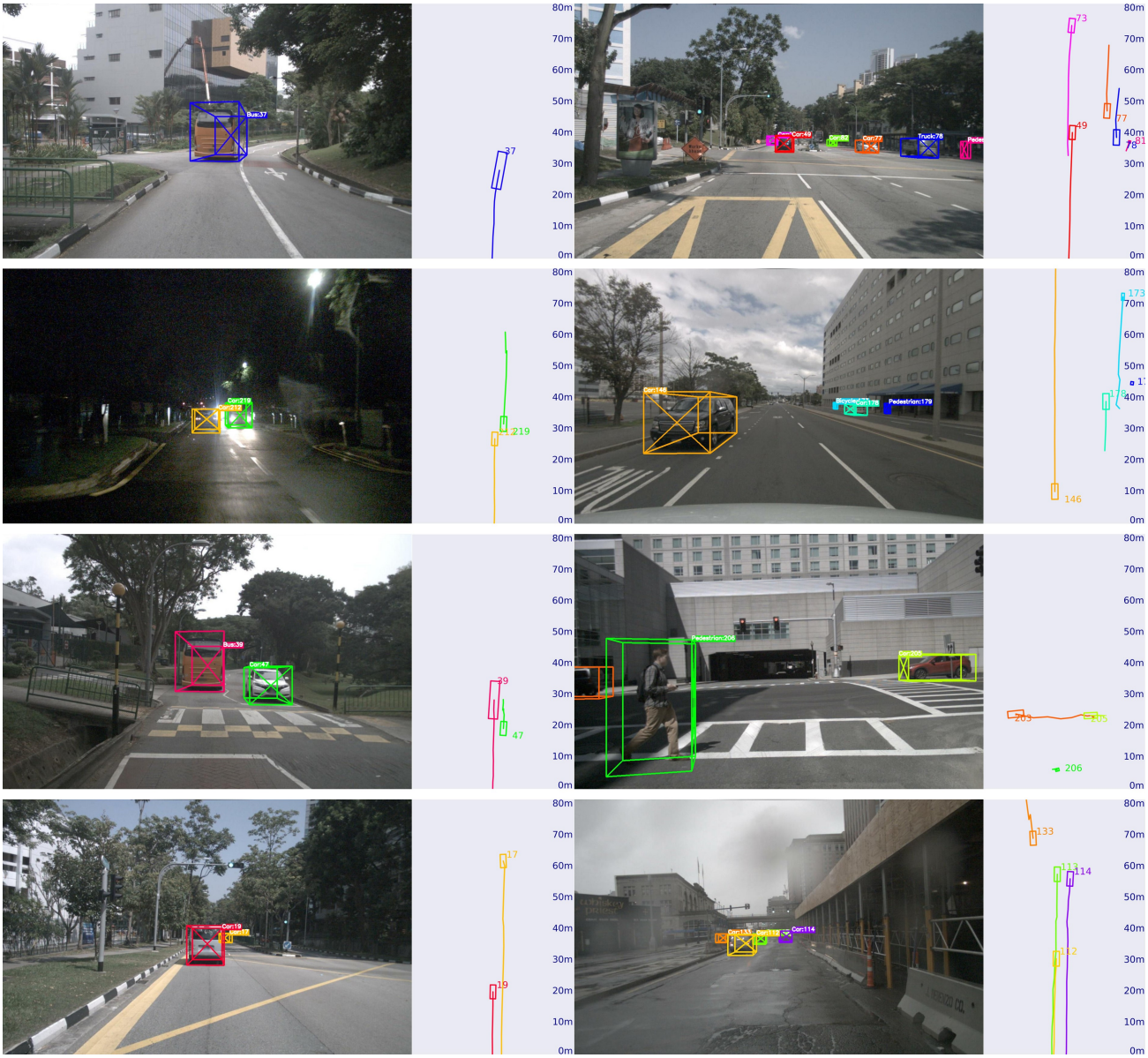

Figure 1. **Qualitative results on the nuScenes validation set.** We plot the tracking results of our MoMA-M3T based on the image view (left) and the bird's eye view (right) with the 15 historical frames on the BEV plane, in which different colors denote different tracklets.

be observed that our method can track different types of objects across various scenarios.

Failure case. We provide a representative failure case in Figure 2. Due to the inaccurate object depth estimation for the yellow box (see the right figure), the tracker cannot associate it with the existing tracklet (#139) since their position distance is too far from each other (more than 10 meters). It thus generates a new identity (#143) for the detection.

References

[1] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008. 1

[2] Mohamed Chaabane, Peter Zhang, Ross Beveridge, and Stephen O'Hara. Deft: Detection embeddings for tracking. In *CVPR Workshops*, 2021. 1

[3] Tobias Fischer, Yung-Hsu Yang, Suryansh Kumar, Min Sun, and Fisher Yu. Cc-3dt: Panoramic 3d object tracking via cross-camera fusion. In *CoRL*, 2022. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1

[5] Hou-Ning Hu, Yung-Hsu Yang, Tobias Fischer, Trevor Darrell, Fisher Yu, and Min Sun. Monocular quasi-dense 3d object tracking. *TPAMI*, 2022. 1

Figure 2. **Representative failure case.** The solid lines and boxes denote the tracking results for timestamps $t-1$ and t, while the dotted one is the mismatched tracklet (#139) in the coordinate at t timestamp. The failure case is caused by the inaccurate observation (yellow box at frame t) from the monocular 3D object detector. The tracker cannot associate the detection with any tracklet (*e.g.*, #139), thus generating a new identity for it (#143). Note that, we only plot a few bounding boxes for better illustrations in this example.

[6] Peixuan Li and Jieyu Jin. Time3d: End-to-end joint monocular 3d object detection and tracking for autonomous driving. In *CVPR*, 2022. 1

[7] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 1

[8] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *ICCV*, 2021. 1

[9] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, , and Justin M. Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *CoRL*, 2021. 1

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

[11] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *CVPR*, 2021. 1

[12] Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In *CVPR Workshops*, 2022. 1

[13] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020. 1