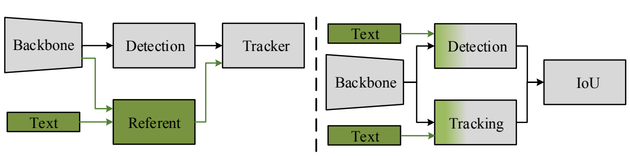
Supplementary Material for Referring Multi-Object Tracking

Dongming Wu[1*‡], Wencheng Han[2*], Tiancai Wang[3], Xingping Dong[4], Xiangyu Zhang[3,5], Jianbing Shen[2†]
[1] Beijing Institute of Technology, [2] SKL-IOTSC, CIS, University of Macau, [3] MEGVII Technology,
[4] School of Computer Science, Wuhan University, [5] Beijing Academy of Artificial Intelligence
{wudongming97, wenchenghan, shenjiangbingcg}@gmail.com, wangtiancai@megvii.com

1. Competitor Details

In this section, we provide more model details about the competitors (introduced in §5.2). The CNN-based counterparts build upon several multi-object tracking (MOT) models, such as FairMOT [5], DeepSORT [3], ByteTrack [4], and CStrack [1], with some crucial modifications on cross-modal learning. In specific, these CNN-based MOT models typically follow a *tracking-by-detection* pattern, which consists of a detector (including backbone and detection head) for single-frame detection and a tracker for cross-frame object association. As shown in Fig. 1(a), we design a referent branch on the visual backbone. It contains our proposed cross-modal fusion module and the detection head from the original MOT model. The cross-modal module fuses visual and linguistic features and provides comprehensive feature representation. The detection head decodes the fused feature maps into object boxes with the same format as the original outputs. During training, we keep the losses of predicting all visible objects. For inference, the default tracker is used to associate cross-frame referent objects. DeepSORT and ByteTrack do not provide a detection model, so we employ the referent results from FairMOT.

In addition to CNN-based methods, we also experiment with a Transformer-based MOT model, TransTrack [2]. We modify it by adding our cross-modal early-fusion module before the encoder layers, as depicted in Fig. 1(b). Both TransTrack and our TransRMOT belong to Transformer-based frameworks. But TransTrack is not an end-to-end model as it uses IoU-matching between a detection model and a tracking model to determine the final referent objects.



(a) CNN-based models (b) Transformer-based models

Figure 1. The original MOT models and our cross-modal modification for CNN-based and Transformer-based framework.

2. Limitation

Fig. 2 visualizes several failure cases from TransRMOT. The first case is that some fine-grained object features (*e.g.*, human gender) are not captured accurately, hindering the detection performance. To avoid this case, the top-down solution (*i.e.*, the detection-then-fusion method) can be jointly explored to focus more on the fine-grained features of object regions. The second case is ID switch problem, which is caused by long-temporal occlusion and degrades the tracking performance. To address this problem, object representation can keep more time for long-term association using a memory mechanism in future work.

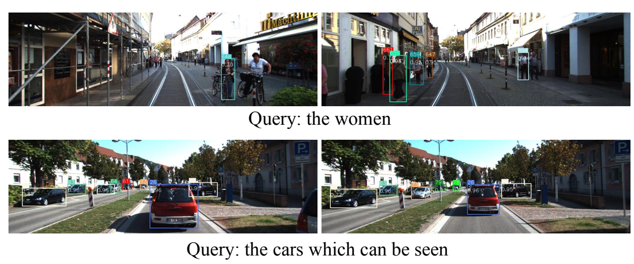


Figure 2. Typical failure cases from TransRMOT.

3. More Qualitative Results

We offer more qualitative results in Fig. 3. As seen, our proposed TransRMOT achieves compelling results under various challenging situations, *e.g.*, multiple objects, object entrance and exit, moving objects, occlusion and *etc*.

References

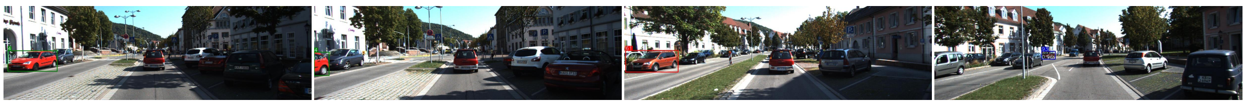
[1] Chao Liang, Zhipeng Zhang, Xue Zhou, Bing Li, Shuyuan Zhu, and Weiming Hu. Rethinking the competition between detection and reid in multiobject tracking. *IEEE TIP*, 2022. 1

[2] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 1

[3] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 1

Query: the cars in front of ours

Query: the left cars which are parking

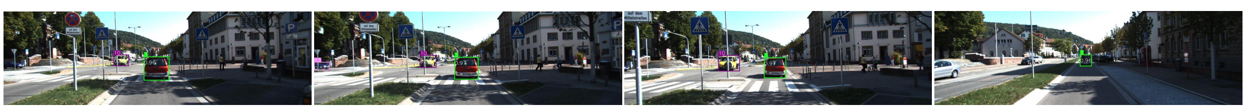


Query: the left cars in red

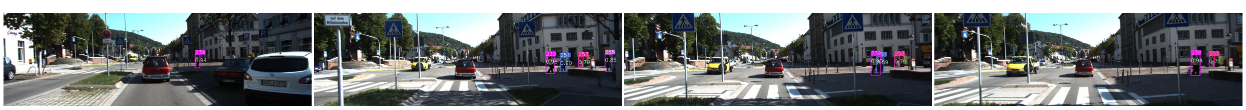
Query: the left cars in black

Query: the left cars in light color

Query: the parking cars



Query: the cars which are moving



Query: the pedestrian

Query: the right pedestrian

Figure 3. More qualitative results on Refer-KITTI.

[4] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022. 1

[5] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 2021. 1