Supplementary Material for Referring Multi-Object Tracking

Dongming Wu\textsuperscript{1*}, Wencheng Han\textsuperscript{2*}, Tiancai Wang\textsuperscript{3}, Xingping Dong\textsuperscript{4}, Xiangyu Zhang\textsuperscript{3,5}, Jianbing Shen\textsuperscript{2†}

\textsuperscript{1} Beijing Institute of Technology, \textsuperscript{2} SKL-IOTSC, CIS, University of Macau, \textsuperscript{3} MEGVII Technology, \textsuperscript{4} School of Computer Science, Wuhan University, \textsuperscript{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 \cite{liang2022rethinking}, DeepSORT \cite{wojke2017simple}, ByteTrack \cite{sun2020transtrack}, and CStrack \cite{liang2022rethinking}, with some crucial modifications on cross-modal learning. In specific, these CNN-based MOT models typically follow a \textit{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 \cite{sun2020transtrack}. 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.

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.

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

\cite{liang2022rethinking, sun2020transtrack, wojke2017simple}

\textsuperscript{*} Corresponding author.

\textsuperscript{†} Senior author.

\textsuperscript{1} Beijing Institute of Technology, \textsuperscript{2} SKL-IOTSC, CIS, University of Macau, \textsuperscript{3} MEGVII Technology, \textsuperscript{4} School of Computer Science, Wuhan University, \textsuperscript{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 \cite{liang2022rethinking}, DeepSORT \cite{wojke2017simple}, ByteTrack \cite{sun2020transtrack}, and CStrack \cite{liang2022rethinking}, with some crucial modifications on cross-modal learning. In specific, these CNN-based MOT models typically follow a \textit{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 \cite{sun2020transtrack}. 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.

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.

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

\cite{liang2022rethinking, sun2020transtrack, wojke2017simple}
Figure 3. More qualitative results on Refer-KITTI.
