DeconfuseTrack: Dealing with Confusion for Multi-Object Tracking

Supplementary Material

A. More Ablation Studies

Method	DanceTrack-val[43]				MOT17-val[32]
	HOTA↑	IDF1↑	AssA↑	FPS↑	FPS↑
Baseline	52.7	53.0	35.3	**48.8**	**45.6**
Baseline+DDA	56.2	57.0	40.3	44.1	33.0
Baseline+ONMS	53.7	54.0	36.7	46.2	41.5
DeconfuseTrack	**56.8**	**57.3**	**41.1**	39.7	20.7

Table 6. Additional ablation experiments were conducted on both the DanceTrack and MOT17 validation sets. The tracking metrics for the MOT17 validation set have already been provided in Tab. 4.

A.1. Ablation study on DanceTrack

DanceTrack[43] is a dataset consisting of dance videos, where the targets exhibit highly similar appearances and complex nonlinear motion patterns. When using the Kalman filter[22] for motion prediction, there is significant ambiguity due to the characteristics of dataset. Tab. 6 presents the results of our proposed method on the DanceTrack validation set, with all hyperparameters set consistently with the MOT17[32] validation set. It can be observed that our proposed method consistently improves several tracking metrics, demonstrating the adaptability of our approach. Compared to MOT17, DeconfuseTrack exhibits larger improvements across various metrics, with a 5.8% increase in AssA, a 4.3% increase in IDF1, and a 4.1% increase in HOTA. We hypothesize that the increased complexity of target motion in DanceTrack, along with the higher ambiguity in motion prediction, allows us to mitigate some of the confusion even when targets have similar appearances, leveraging appearance cues.

A.2. Speed

Tab. 6 also presents the runtime of our proposed method (GPU: NVIDIA RTX 4090). The additional computational overhead compared to the baseline mainly arises from the inclusion of the appearance model, the incorporation of more unreliable detections through ONMS, and the various disambiguation modules. It is important to note that we have not implemented any specific optimizations for CPU computations at the moment. It can be observed that our method can achieve real-time performance. When the target density is high, we propose grouping trajectories and detections based on their positions to reduce the computational load of each disambiguation module.

Method	HOTA↑	IDF1↑	MOTA↑	AssA↑
DDM+ONMS	+1.0	+1.7	+0.0	+2.1
TDM+ONMS	+0.4	+1.0	-0.1	+1.0
ADM+ONMS	+0.1	+0.1	+0.0	+0.2

Table 7. The results of applying ONMS separately to DDM, TDM, and ADM on the MOT17 validation set are shown. To emphasize the variations, we only present the differences in metrics, while the original values for each module can be referred to in Tab. 3.

A.3. Impact of ONMS

In order to study the gain of ONMS on each individual disambiguation module, we have compared the gains achieved by ONMS in Tab. 7. ONMS demonstrates the most significant improvement in DDM (HOTA +1.0, AssA +2.1) since it provides more unreliable detections that can be utilized by DDM. TDM exhibits a moderate gain (HOTA +0.4, AssA +1.0), while the improvement in ADM is minimal (HOTA +0.1, AssA +0.2). This is because the additional unreliable detections pose challenges in forming easily confusable association pairs in ADM.