GRAE-3DMOT: Geometry Relation-Aware Encoder for Online 3D Multi-Object Tracking

Supplementary Material

S-1. Edge representation in graph construction

We conduct additional experiments to further analyze the edge representation methods in graph construction, which impact feature aggregation. In Table S-1, we conduct the comprehensive ablation study to compare the impacts of the threshold d_S and the edge definition $\mathbf{e}_{ij}^{(t)}$ in the spatial relation graph. From the results at $d_S = 5$, we observe that the configuration of $\mathbf{g}_j^{(t)} - \mathbf{g}_i^{(t)}$ achieves the best AMOTA and AMOTP. This observation is consistent even at $d_S = \infty$. Furthermore, in general, $d_S = \infty$ yields the performance improvement in AMOTA and AMOTP compared to $d_S = 5$ for all edge configurations. These results demonstrate that the proposed edge representation, which considers relative geometry information ($\mathbf{e}_{ij}^{(t)} = \mathbf{g}_j^{(t)} - \mathbf{g}_i^{(t)}$) and leverages extensive geometry relations ($d_S = \infty$) for feature aggregation, is effective for accurate 3D MOT.

d_S	$\mathbf{e}_{ij}^{(t)}$	AMOTA↑	AMOTP↓	MOTA↑	IDS↓	FRAG↓
5	$\mathbf{g}_j^t$	0.720	0.495	0.620	416	213
	$\mathbf{f}_j^t$	0.719	0.502	0.621	477	**199**
	$\mathbf{f}_j^t - \mathbf{f}_i^t$	0.725	0.498	0.623	381	296
	$\mathbf{g}_j^t - \mathbf{g}_i^t$	0.732	0.495	0.632	200	309
∞	$\mathbf{g}_j^t$	0.725	0.494	0.630	277	359
	$\mathbf{f}_j^t$	0.724	0.498	0.627	370	241
	$\mathbf{f}_j^t - \mathbf{f}_i^t$	0.729	0.493	0.631	208	357
	$\mathbf{g}_j^t - \mathbf{g}_i^t$	**0.737**	**0.488**	**0.636**	**150**	315

Table S-1. Ablation study for edge representation in spatial relation graph construction.

S-2. Ablation study for hyperparameters

Track age. Table S-2 shows the 3DMOT performance according to track age threshold λ. We observe that the performance is degraded as λ decreases. For instance, ID switches (IDS) increases with fewer λ, since issues of occlusions and missed detection cannot be addressed. We select $\lambda = 12$, which yields the best AMOTA and AMOTP performance.

λ	AMOTA↑	AMOTP↓	MOTA↑	IDS↓	FRAG↓
3	0.719	0.528	0.627	288	416
6	0.727	0.499	0.633	176	322
9	0.732	0.489	0.635	153	**314**
12	**0.737**	**0.488**	0.636	**150**	315
15	0.736	0.489	**0.637**	**150**	319

Table S-2. Ablation study for hyperparameter λ.

Model dimension. Table S-3 lists the performance according to the feature dimension C. In general, the performance improves as C increase, while efficiency decrease. We pick $C = 128$, which yields the best performance while satisfying real-time processing requirements.

C	AMOTA	AMOTP	FPS↑	Parameter (MB)↓	FLOPs (MB)↓
32	0.733	0.489	**57.30**	**0.04**	**0.06**
64	0.735	0.490	57.21	0.14	0.24
128	**0.737**	**0.488**	56.19	0.57	0.97
256	0.736	**0.488**	49.15	2.26	3.86

Table S-3. Ablation study for feature dimension C.

S-3. Ablation study for track feature

In the proposed GRAE-3DMOT, the track feature is subsequently updated for aggregation with the spatiotemporal relation-aware features at next frame. Table S-4 compares the performance with and without the track feature in association score computation. We see that the track feature significantly improves the AMOTA and AMOTP performance based on memory information from previous frames.

	AMOTA↑	AMOTP↓	MOTA↑	IDS↓	FRAG↓
w/o track feature	0.731	0.494	0.629	193	**264**
with track feature	**0.737**	**0.488**	**0.636**	**150**	315

Table S-4. Ablation study for track feature.

S-4. MLP detail

The MLP consists of a linear layer, an activation function (ReLU), and another linear layer.

S-5. Experiments on Waymo

Waymo [2] consists of 798 training sequences, 202 validation sequences, and 150 test sequences. Waymo provides LiDAR sensors and multiview cameras. The annotations contain 3 common classes: vehicle, pedestrian, and cyclist. We train GRAE-3DMOT using the training sequences and evaluate the trained GRAE-3DMOT on the validation set.

Table S-5 compares the 3D MOT performance of GRAE-3DMOT with 3DMOTFormer [1] on the Waymo validation dataset. To obtain detections, we use the same detector, CasA [3], for both GRAE-3DMOT and 3DMOTFormer. The proposed GRAE-3DMOT outperforms 3DMOTFormer only except MOTP for the Vehicle category.

	MOTA↑			MOTP↓		
	Vehicle	Pedestrian	Cyclist	Vehicle	Pedestrian	Cyclist
3DMOTFormer	45.70	60.51	57.76	**16.33**	30.71	**25.87**
GRAE-3DMOT	**58.23**	**62.39**	**57.79**	16.59	**30.43**	**25.87**

Table S-5. Results on Waymo [2] validation set using CasA [3] detections. The best results are boldfaced.

References

[1] Shuxiao Ding, Eike Rehder, Lukas Schneider, Marius Cordts, and Juergen Gall. 3dmotformer: Graph transformer for online 3d multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9784–9794, 2023. 1

[2] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 2

[3] Hai Wu, Jinhao Deng, Chenglu Wen, Xin Li, Cheng Wang, and Jonathan Li. Casa: A cascade attention network for 3-d object detection from lidar point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 1, 2