Heterogeneous Diversity Driven Active Learning for Multi-Object Tracking (Supplementary Materials)

Rui Li[1,2,*], Baopeng Zhang[1], Jun Liu[2], Wei Liu[1], Jian Zhao[3,4,5], Zhu Teng[1,†]

[1]Beijing Jiaotong University [2]Singapore University of Technology and Design
[3]Institute of North Electronic Equipment [4]Peng Cheng Laboratory [5]Intelligent Game and Decision Laboratory

In this supplementary material, we provide illustrations that show more details on the design of our HD-AMOT model, present additional experiments of the hyper-parameter study, and show more quantitative results of active learning for multi-object tracking.

1. HD-AMOT Details

Diversified Informative Representation Details. In this section, we provide illustrations that show more details on how feature spaces s_{set} and s_{frame} are built according to the output of the tracker f_t. Fig. 1(A) shows how to build the feature space S_{set} of set-level discrepancy and Fig. 1(B) how to build the feature space S_{frame} of frame-level diversity. To facilitate the calculation of maximum mean discrepancy (MMD) and cosine similarity in diversified informative representation learning, we transform the heterogeneous clues obtained by the tracker f_t to vector features in feature spaces.

Specifically, when learning the set-level discrepancy, we perform global average pooling and max pooling on the global semantics $\boldsymbol{F}^f$ and the local semantics $\{\boldsymbol{F}^{f_1}, \boldsymbol{F}^{f_2}, ..., \boldsymbol{F}^{f_K}\}$ respectively. Then the results of average pooling and max pooling are concatenated, followed by flattening to generate the semantic feature vectors. For the global spatial topology $\boldsymbol{G}^h$ and local spatial topology $\{\boldsymbol{G}^{h_1}, \boldsymbol{G}^{h_2}, ..., \boldsymbol{G}^{h_K}\}$, they are also flattened to generate topology feature vectors. Finally, these semantic feature vectors and topology feature vectors make up our feature space S_{set}, which is used to calculate the maximum mean discrepancy between the labeled subset $\boldsymbol{X}_t^L$ and unlabeled subset $\boldsymbol{X}_t^U$ to generate the state representation s_t of our design MDP. In frame-level diversity learning, the global semantics $\boldsymbol{F}^f$ and global spatial topology $\boldsymbol{G}^h$ are flattened in the same way to get the corresponding feature vectors. In addition, the object scale matrix $\boldsymbol{G}^b$ is flattened to generate two scale vectors. These vectors of global semantics, global spatial topology, and object scale constitute the feature space S_{frame} to generate the action a_t of each unlabeled frame for our design MDP, where the cosine similarity distribution is recorded to learn the histogram-based representation.

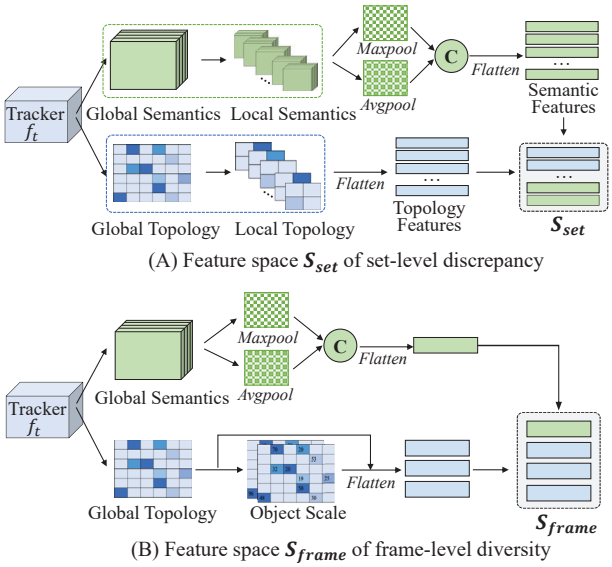

Figure 1. The construction process of S_{set} and S_{frame}.

Informative Frame Selection Network Details. The detailed illustration of the lightweight IFSN is shown in Fig. 2. Firstly, s_t and o_t are fed into a linear layer with ReLU activation to generate a 32-D feature. Then, s_t, a_t, and the 32-D cooperation feature are concatenated and passed through two linear layers with ReLU activation to obtain the 128-D feature. Finally, a final linear layer with ReLU activation is applied to obtain the score of an unlabeled frame. Regarding the fixed-length compact representation o_t obtained in multi-frame cooperation, the illustration of its generation is

*Work done when Rui Li was an visiting student of SUTD.
†Zhu Teng is the corresponding author (zteng@bjtu.edu.cn).

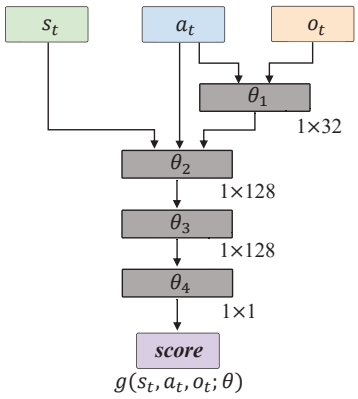

Figure 2. The architecture of the designed lightweight IFSN.

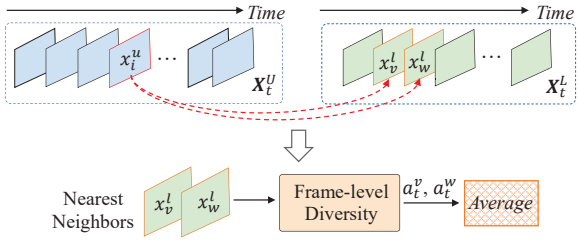

Figure 3. The illustration of multi-frame cooperation.

shown in Fig. 3.

2. Additional Experiments

In this section, we analyze the influence of hyper-parameters of our HD-AMOT framework: the number of local spatial topologies (local semantic features) and the partition ratio of the reward subset $\boldsymbol{X}^R/\boldsymbol{X}$. In addition, MOT15 [2], MOT17 [3], MOT20 [1], and Dancetrack [4] datasets are used for complex multi-pedestrian tracking. To verify the generalization of our HD-AMOT, the performance evaluation of multi-car tracking is conducted on BDD100K [5].

Analysis of Different Local Numbers. We investigate the performance of our HD-AMOT framework under different numbers of local spatial topologies (local semantic features) and present the results in Tab. 1. When the number of local parts is 0, it means that the HD-AMOT model only uses global features to compute the set-level discrepancy. Remarkably, an increase in MOTA and IDF1 from 0 local parts to 9 local parts can be observed, which demonstrates that local parts bring greater improvement. We also noted that the upward tendency of MOTA and IDF1 stagnates around 9 local parts, and it shows a downward tendency after that. Hence, we use 9 local parts in the set-level discrepancy learning. Moreover, we observe that the utilization of local parts has a great influence on IDF1 but a slight influence on MOTA, which indicates that the local

Table 1. Ablation study of different numbers of local parts on MOTA and IDF1 metrics.

Number	MOTA (%)	IDF1 (%)
K = 0	62.9	64.6
K = 4	63.2	64.5
K = 9	**63.2**	**66.5**
K = 16	63.0	65.7
K = 25	63.0	65.1

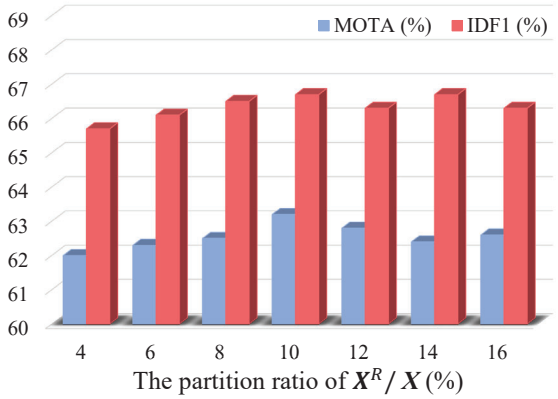

Figure 4. The histograms reflecting the influence of different partition ratios $\boldsymbol{X}^R/\boldsymbol{X}$ on MOTA and IDF1 metrics.

Table 2. Active learning results with a 20% labeled budget for one-stage multi-car tracking.

Sampling Method	MOTA (%)	IDF1 (%)
Random Sampling	25.63	32.95
Uniform Sampling	25.81	33.44
Our HD-AMOT	**26.96**	**34.12**

parts mainly bring significant gain to the object association of one-stage MOT.

Analysis of Different Partition Ratios $\boldsymbol{X}^R/\boldsymbol{X}$. We also look into the performance of our proposed HD-AMOT framework under different partition ratios of the reward subset $\boldsymbol{X}^R$ as shown in Fig. 4. Our HD-AMOT obtains the best MOTA and IDF1 when the partition ratio $\boldsymbol{X}^R/\boldsymbol{X} = 10$. Considering that in one-stage MOT, MOTA is a comprehensive metric to evaluate the detection performance and IDF1 estimates the object association ability of a tracker, we use $\boldsymbol{X}^R/\boldsymbol{X} = 10$ in this paper to obtain the best detection and association performance.

Active Learning for Multi-car Tracking. To verify the generalization of our proposed HD-AMOT, the performance evaluation of multi-car tracking is further conducted on BDD100K, as described in Tab. 2, where ours is su-

perior to other sampling methods. Notably, pedestrians and cars are typical representatives of non-rigid objects and rigid objects respectively. The superior performance of our proposed HD-AMOT on pedestrians and cars illustrates the ability of our method to be extended to other classes.

References

[1] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003*, 2020.

[2] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942*, 2015.

[3] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv:1603.00831*, 2016.

[4] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, 2022.

[5] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020.