AttTrack: Online Deep Attention Transfer for Multi-object Tracking

Supplementary Material

Keivan Nalaie
McMaster University
nalaiek@mcmaster.ca

Rong Zheng
McMaster University
rzheng@mcmaster.ca

1. Experiments on HRNet-W18[1]

In Table 1, the results of Teacher-only and Student-only baselines on the MOT15 and MOT17 datasets are reported. The performance gap between large and small models in MOT17 is larger than in the MOT15 dataset. The large model's capability for locating and characterizing small objects in high-resolution images is the primary cause.

Table 1: Performance of Teacher-only and Student-only on MOT15 and MOT17 datasets.

Baseline	Dataset	MOTA (%)	FPS
Teacher-only	MOT15	64.50	12.47
	MOT17	64.20	12.30
Student-only	MOT15	61.30	19.99
	MOT17	55.60	19.49

Table 2 shows AttTrack's performance with EFM settings. The MOTA gap between $K = 2$ and $K = 6$ on MOT15 is only 1.60% while this metric for MOT17 is 3.40%. We believe these performance differences appear to be related to the corresponding teacher and student baseline models presented in Table 1. Our attention transfer approach can improve the tracking performance of the student model by efficiently extracting and transforming attention to the student model. We substitute Layerwise and Naive-Mix for AttTrack as two different baselines to assess the effectiveness of our attention transmission approach. As shown in Table2, AttTrack outperforms Layerwise and Naive-Mix approaches.

Table 2: Comparing AttTrack with Layerwise and Naive-Mix on HRNet-W18 architecture

Dataset	K	AttTrack-EFM		Layerwise		Naive-Mix	
		MOTA	FPS	MOTA	FPS	MOTA	FPS
MOT15	2	**63.50**	15.58	62.70	15.47	62.60	15.59
	3	62.50	16.58	61.50	16.89	61.60	17.00
	4	62.10	17.64	61.50	17.56	61.10	17.77
	5	62.00	17.88	61.60	18.00	60.90	18.17
	6	61.90	18.04	61.30	18.50	61.10	18.52
MOT17	2	**60.10**	14.84	59.90	15.14	58.90	15.16
	3	58.10	15.93	57.30	16.28	56.20	16.30
	4	57.50	16.65	56.90	17.09	56.10	17.00
	5	56.90	17.16	56.00	17.51	55.70	17.36
	6	56.70	17.40	56.70	17.91	55.80	17.50

References

[1] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware Representation Learning for Bottom-up Human Pose Estimation. In *CVPR*, pages 5386–5395, 2020.