Segment as Points for Efficient Online Multi-Object Tracking and Segmentation

Supplementary Material

Paper ID 1059

Abstract. Demo video of our PointTrack[1] is provided alongside this PDF file. It is worth noting that PointTrack achieves superior MOTS results with a near real-time speed (22 FPS).
In this supplementary material, we describe more details about PointTrack in three aspects. Firstly, we provide a comparison of car density between our APOLLO MOTS and KITTI MOTS, and a sample video from APOLLO MOTS. Secondly, we show ablation studies concerning the compact of scale factor (k), the point weighting layer, and the number of selected points in computing the instance similarity. Lastly, we describe the choice of hyper-parameters in fine-tuning SpatialEmbedding.

1 APOLLO MOTS

A demo video of APOLLO MOTS named *anno_sample_APOLLO_MOTS.avi* is provided alongside this PDF file, where both the segmentation task and the tracking task are highly challenging. Moreover, the comparison of car density between APOLLO MOTS and KITTI MOTS is shown in Fig. 1. Although our APOLLO MOTS has a similar number of frames with that of KITTI MOTS, we has two times more tracks and car annotations.

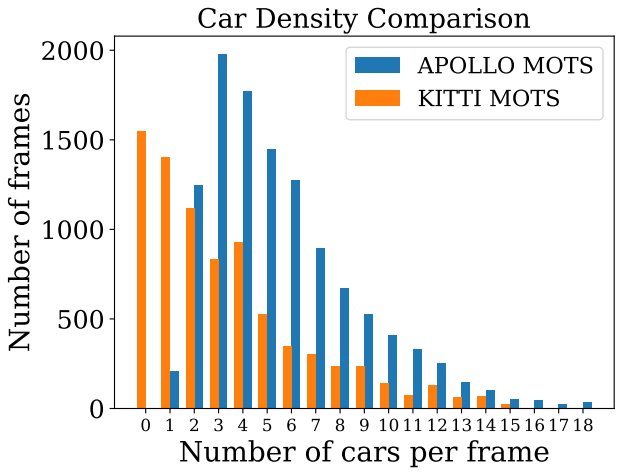

Fig. 1. Comparison of car density between APOLLO MOTS and KITTI MOTS.

[1] named *demo_video_KITTI_MOTS.mp4*

2 Ablation Study

	Cars		Pedestrians	
scale factor(k)	sMOTSA	MOTSA	sMOTSA	MOTSA
0.0	85.36	94.78	61.95	76.93
0.1	85.41	94.83	62.07	77.05
0.2	**85.51**	**94.93**	**62.37**	**77.35**
0.3	85.37	94.79	62.04	77.02

Table 1. Experiments on impact of different scale factors k on the performance of PointTrack.

Impact of different scale factor. When k is set to zero, we extract environment embeddings from the pixels which are inside the bounding box but outside of the segment. Larger k which brings a larger environment area might increase the information capacity of environment embeddings. However, at the same time, as the number of randomly sampled points are fixed, larger environment area results in that informative points become less possible to be selected. As shown in Table 1, we find that PointTrack performs best when k is set to 0.2. Thus, by default, we set k to 0.2 in experiments.

				Cars		Pedestrians	
P_W	M_F	M_E	M_I	sMOTSA	MOTSA	sMOTSA	MOTSA
√	√	√	√	**85.51**	**94.93**	**62.37**	**77.35**
x				85.37	94.79	62.04	77.02
	x			83.59	93.01	61.27	76.25
		x		85.30	94.72	61.98	76.96
			x	85.33	94.76	62.31	77.29

Table 2. Experiments on impact of the point weighting layer, M_F, M_E, and the mask IOU.

Impact of the point weighting layer, M_F, M_E, and the mask IOU. We remove the point weighting layer (P_W), the foreground embeddings M_F, the environment embeddings M_E, and the mask IOU (M_I) in turn to examine their impacts on performance. When we remove the mask IOU, we set α to zero in Eq. (5) in the paper. As shown in Table 2, when the foreground embeddings M_F is removed, the performance drops a lot, demonstrating that the foreground point cloud in the segment area matters most in the instance association. By contrast, when the mask IOU is removed, the performance drop is minimal, especially for Pedestrians. Therefore, for instances with rigid shapes, considering the mask IOU in computing similarity is more beneficial than instances with non-rigid shapes like Pedestrians.

N_F	N_E	Cars		Pedestrians		speed(ms)
		sMOTSA	MOTSA	sMOTSA	MOTSA	
500	250	85.45	94.87	62.01	76.99	**7.3**
1000	500	85.51	94.93	62.37	77.35	7.9
1000	1000	85.51	94.93	62.37	77.35	8.3
2000	1000	**85.53**	**94.96**	**62.38**	**77.37**	9.4

Table 3. Experiments on impact of the number of points selected for feature extraction. The speed of instance association is measured in milliseconds per frame.

Impact of N_F and N_E. More selected points are beneficial for feature extraction. However, for small instances in the image plane where segments only occupy hundreds of pixels, selecting too many points leads to no performance gains and at the same time introduces heavier computations in instance association. As shown in Table 3, considering the trade-off between efficiency and performance, we set N_F and N_E to 1000 and 500, respectively.

3 Details about SpatialEmbedding

Our proposed seed consistency loss is combined with the original seed loss for foreground pixels with the same foreground weight as described in the paper of SpatialEmbedding [24]. We adopt the elliptical margin rather than circular margin for both cars and pedestrians.

When we fine-tuning SpatialEmbedding on KITTI MOTS, we select different foreground weights for cars and Pedestrians. On the GitHub page of SpatialEmbedding, the authors explain that the foreground weight parameter is essential for the instance segmentation performance. Empirically we find PointTrack achieves the best performance under the following settings. For cars, the foreground weight is set to 200. For pedestrians, the foreground weight is set to 50. Afterward, we fine-tune it on KITTI MOTS with our proposed seed consistency loss for 50 epochs at a learning rate of $5 \cdot 10^{-6}$.