Supplementary Material: Joint Spatial-Temporal Optimization for Stereo 3D Object Tracking

Peiliang Li, Jieqi Shi, and Shaojie Shen
The Hong Kong University of Science and Technology
pliap@connect.ust.hk, jshias@connect.ust.hk, eeshaojie@ust.hk

2D Tracking Evaluation and Analysis. Although this paper mainly focuses on the 3D object tracking, we submit our results to the KITTI 2D tracking test server to provide readers a complete reference to our 3D tracking system. As we introduced in Sect 3.1, to make our framework simple and efficient, we extend the joint stereo proposal strategy in [3] to the sequential images, which enables us simultaneously detect and associate 2D objects without additional pair-wise similarity computation. Though simple, our 2D tracker demonstrates good tracking performance compared to recent state-of-the-art 2D tracking methods as shown in Table. 1.

An interesting phenomenon is that our 2D tracker produces lowest False Positives (FP) and higher False Negatives (FN) compared to [8, 2, 5], which can be explained by the characteristic of the paired proposal. Since only the anchor which overlaps with the union area of the sequential 2D object box will be treated as a foreground proposal, i.e., the "alarm threshold" for positive samples is increased, which significantly reduces the false positive rate. Similarly, due to the variant location (nearby large objects, fast motion, etc) of the object on adjacent images, a set of predefined anchors may miss covering some distantly located pairs, which can be potentially overcome with the help of recent anchor-free 2D detection approaches [7]. Although slightly underperform [2] in 2D tracking, we show significant better 3D tracking performance benefit from our joint spatial-temporal optimization. Employing additional object similarity calculation or exploring anchor-free based paired proposal may further boost our association performance, while outside the main scope of this work.

More Qualitative Examples. We visualize more qualitative examples on KITTI, Argoverse Tracking, and 3D pedestrian tracking in Fig. 1 2 3 respectively, where the relative trajectories on the bird's eye view are also showcased.

Method	MOTA ↑	MOTP ↑	MT ↑	ML ↓	# FP ↓	# FN ↓
mmMOT [8]	**84.77**	85.21	73.23	**2.77**	711	4243
Joint-Tracking [2]	84.52	85.64	**73.38**	2.77	705	**4242**
MOTBeyondPixels [5]	84.24	**85.73**	73.23	2.77	705	4247
JCSTD [6]	80.57	81.81	56.77	7.38	405	6217
3D-CNN/PMBM [4]	80.39	81.26	62.77	6.15	1007	5616
FAMNet [1]	77.08	78.79	51.38	8.92	760	6998
Ours (ST-3D)	82.64	83.83	61.69	7.23	**234**	5366

Table 1. **2D tracking results on the KITTI test set.** We mainly list recently published methods (whether they have the 3D tracking ability or not) for reference.

References

[1] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1

[2] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krahenbuhl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. In *The IEEE International Conference on Computer Vision (ICCV)*. 1

[3] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019. 1

[4] Samuel Scheidegger, Joachim Benjaminsson, Emil Rosenberg, Amrit Krishnan, and Karl Granström. Mono-camera 3d multi-object tracking using deep learning detections and pmbm filtering. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 433–440. IEEE, 2018. 1

[5] Sarthak Sharma, Junaid Ahmed Ansari, J Krishna Murthy, and K Madhava Krishna. Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3508–3515. IEEE, 2018. 1

[6] Wei Tian and et al. Online multi-object tracking using joint domain information in traffic scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 2019. 1

[7] Zhi Tian and et al. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*, 2019. 1

[8] Wenwei Zhang, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, and Chen Change Loy. Robust multi-modality multi-object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2365–2374, 2019. 1

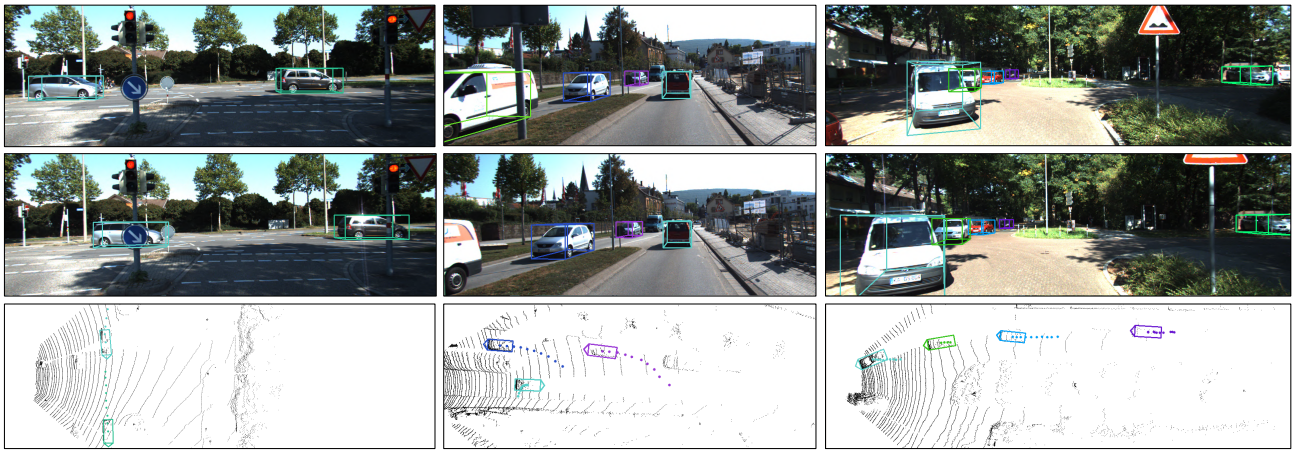

Figure 1. **More qualitative results on the KITTI dataset.** Note that the relative trajectories of 3D objects respecting to the ego-camera are visualized on the bird's eye view image.

Figure 2. **qualitative results on the Argoverse Tracking dataset,** where the stereo images are recored in 5 fps with small FOV cameras.

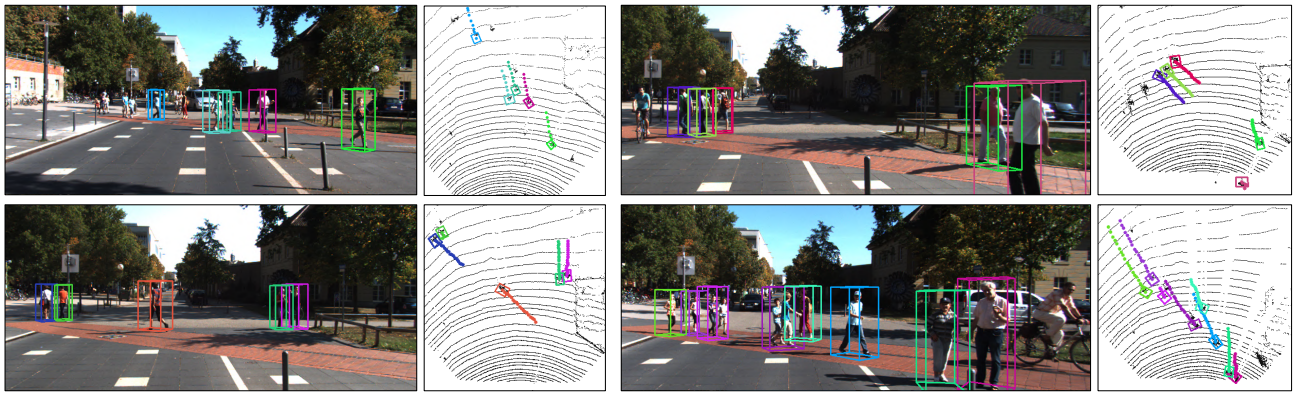

Figure 3. **More qualitative results of the 3D pedestrian tracking on the KITTI dataset.** Each bird's eye view image corresponds to its left RGB image.