Supplementary Material

Anonymous ECCV submission

Paper ID 1811

1 Introduction

In this supplementary materials, we firstly show the detailed architecture of StackNet. In Section 2, additional results of our monocular depth estimation network on the test split [1] of KITTI Raw [2] and also on the Cityscapes dataset [3] using the model trained on KITTI to show the generalization ability of our model. In Section 3, we provide more results of DVSO on the KITTI odometry benchmark. Still, we show the generalization ability of DVSO on the Cityscapes *Frankfurt* sequence.

2 Monocular Depth Estimation

The detailed architecture of StackNet is shown in Figure 1.

2.1 KITTI

In Figure 2, we show more qualitative results of our model on the KITTI Eigen test split [1]. In comparison, we also demonstrate the results predicted by the model of Godard et al. [4]. Note that our model is semi-supervisedly trained on KITTI, while the model of Godard et al [4] is self-supervisedly trained on the Cityscapes dataset and the KITTI dataset. The ground truth is interpolated for better visualization.

2.2 Cityscapes

In Figure 3, we show more qualitative results on the Citysacpes dataset. Note that both our model and the model of Godard et al. [4] are trained on KITTI only.

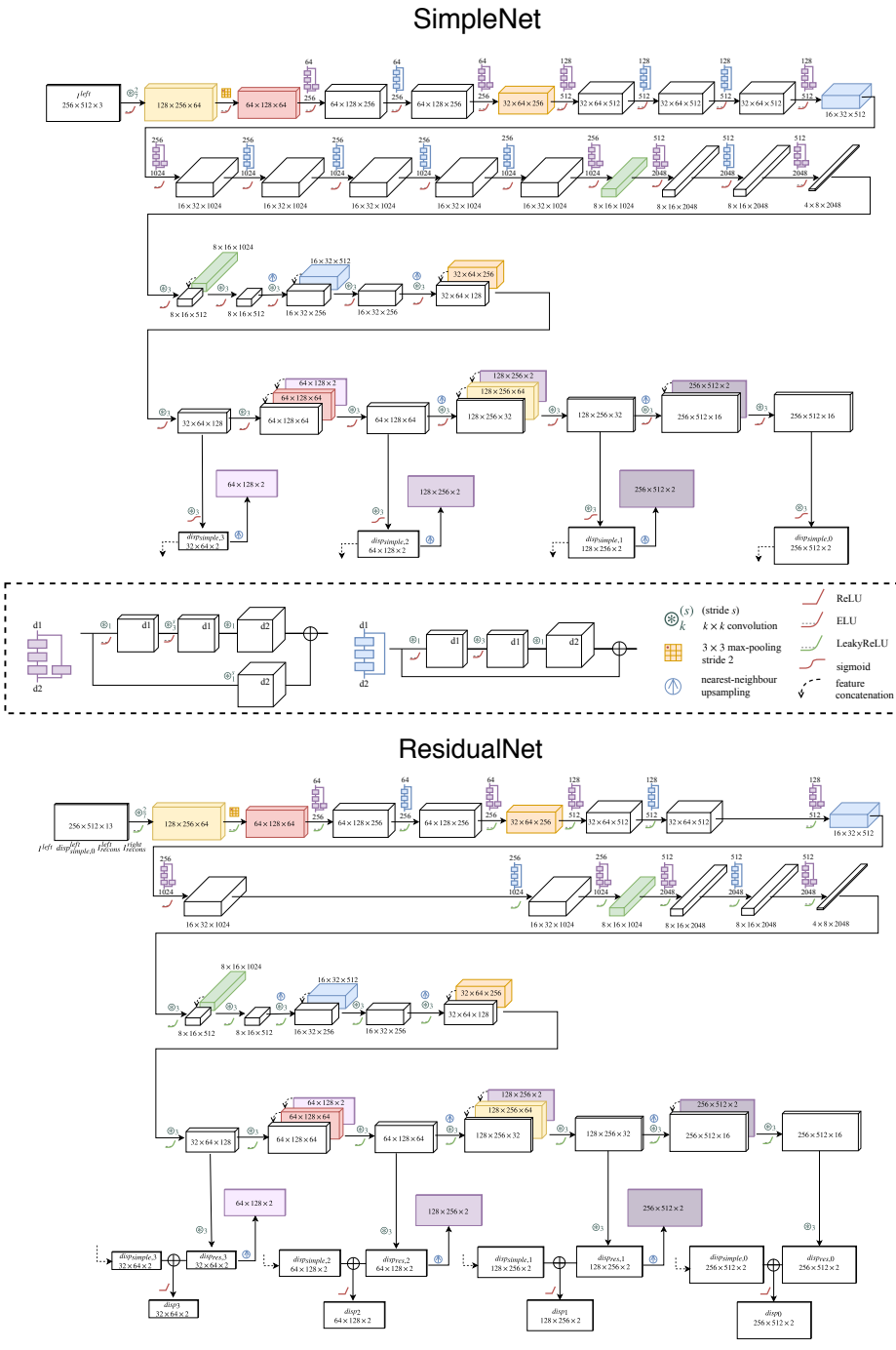


Fig. 1: Detailed network architecture of StackNet.

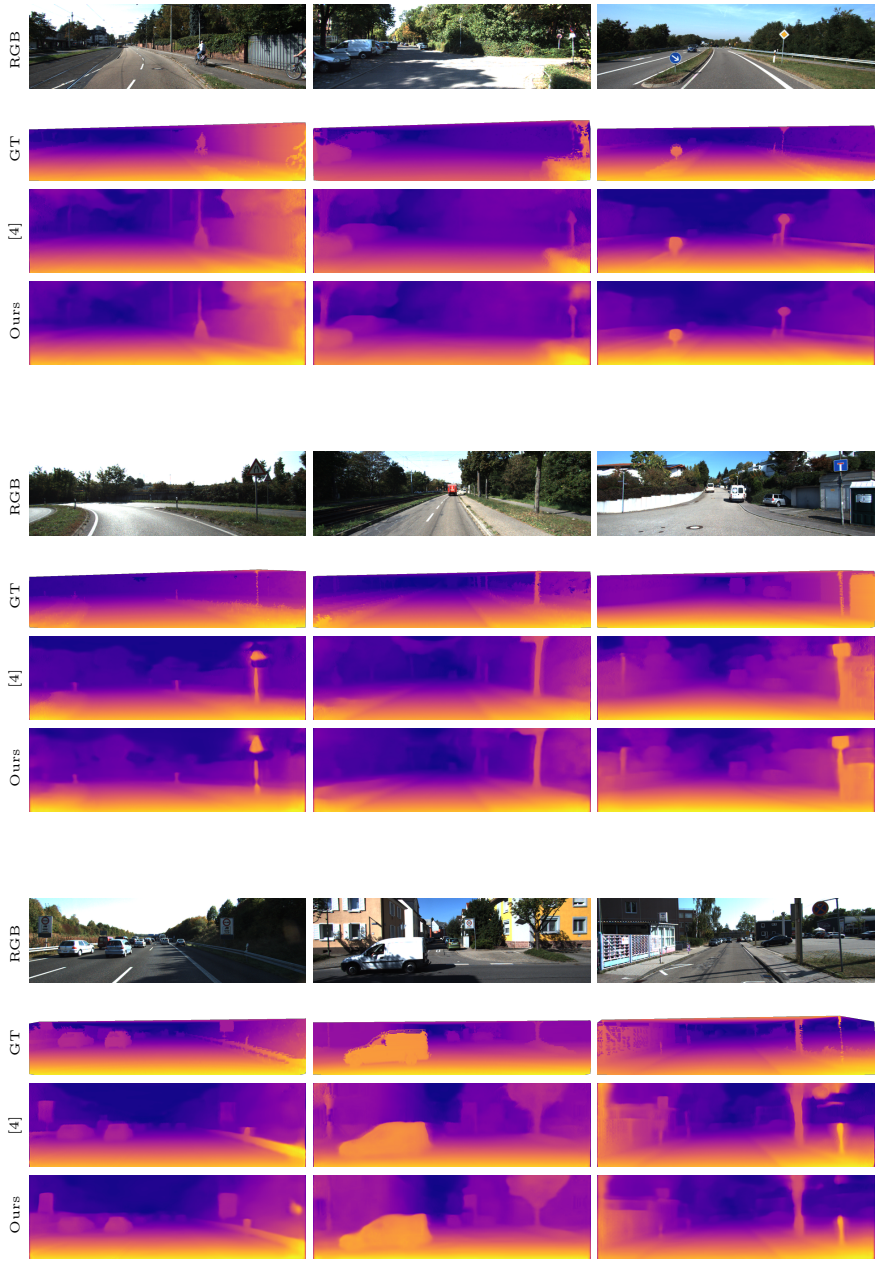


Fig. 2: KITTI. The right-bottom group shows that our method is not able to recover the round traffic sign on the upper right part of the image.

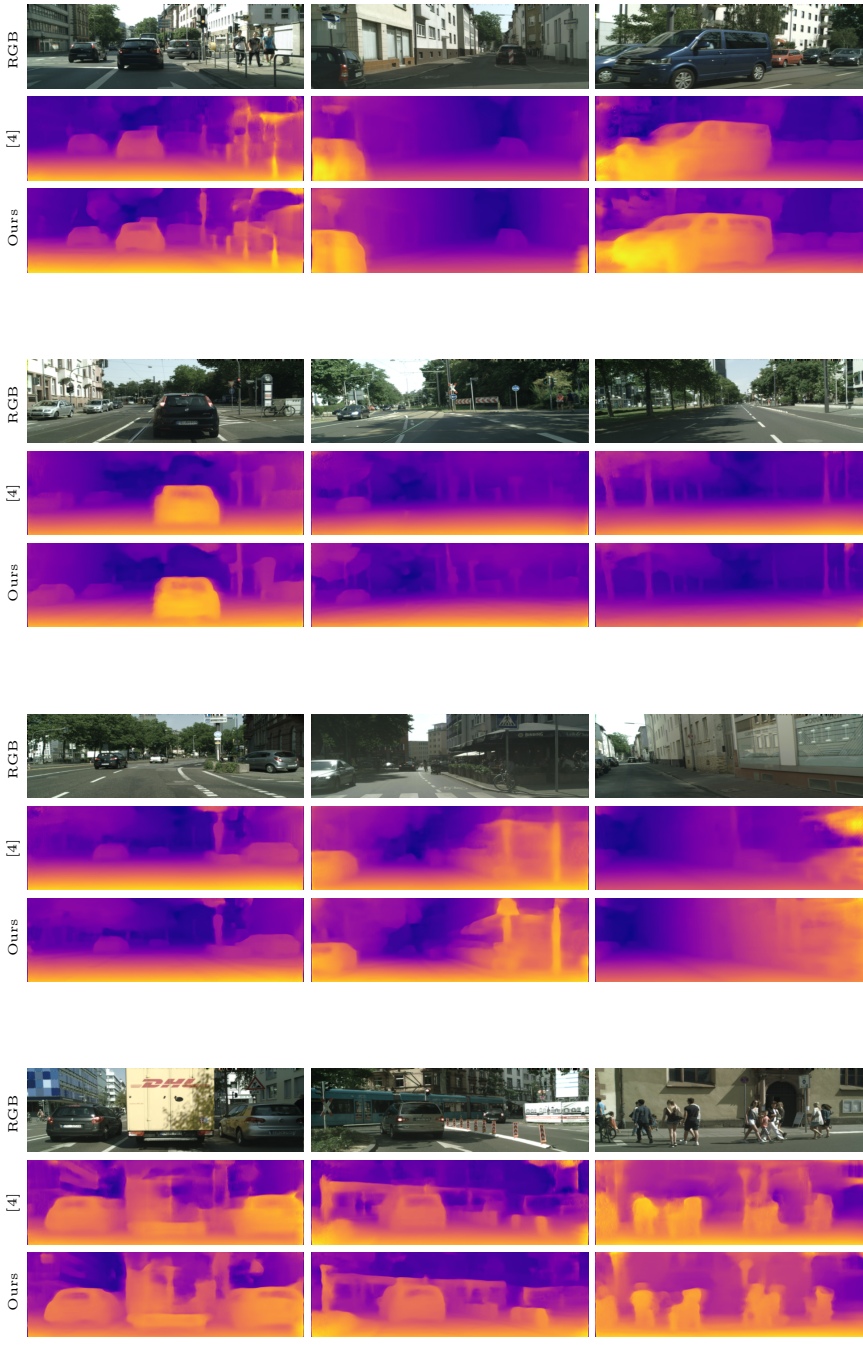


Fig. 3: Cityscapes. The last row shows failure cases where both methods are not able to accurately recover the depth of the van, the train and the various pedestrians.

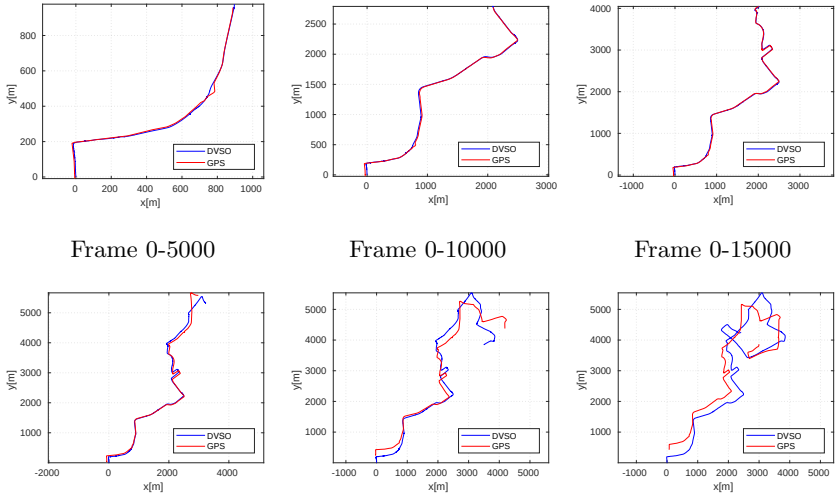
3 Monocular Visual Odometry

3.1 KITTI

In Figures 5 to 7 , we show all the estimated trajectories in the training set (sequences 00-10) of the KITTI odometry benchmark compared with other state-of-the-art monocular [5,6] and stereo [7,8] visual odometry methods. We also show the full evaluation results on the test set in Figure 8 and Figure 9[1].

3.2 Cityscapes

We show the results of DVSO on the Cityscapes *Frankfurt* sequence from frame 0 to 32000, which covers around 10 kilometers path with dynamic environment. The estimated trajectories with different length fragments are shown in Figure 4. Since the camera parameters between KITTI and Cityscapes Dataset are different, the estimated trajectories are Sim(3) aligned to GPS ground truth for qualitative comparison. As can be seen, DVSO works well from on frames 0-20000. Afterwards, the drift becomes larger.



Frame 0-5000 Frame 0-10000 Frame 0-15000

Frame 0-20000 Frame 0-25000 Frame 0-32000

Fig. 4: Estimated trajectories in Cityscapes *Frankfurt*. The estimated trajectories are Sim(3) aligned to GPS ground truth. DVSO works on 0-20000 frames, while the drift becomes larger afterwards.

[1] One can also access the results from KITTI website

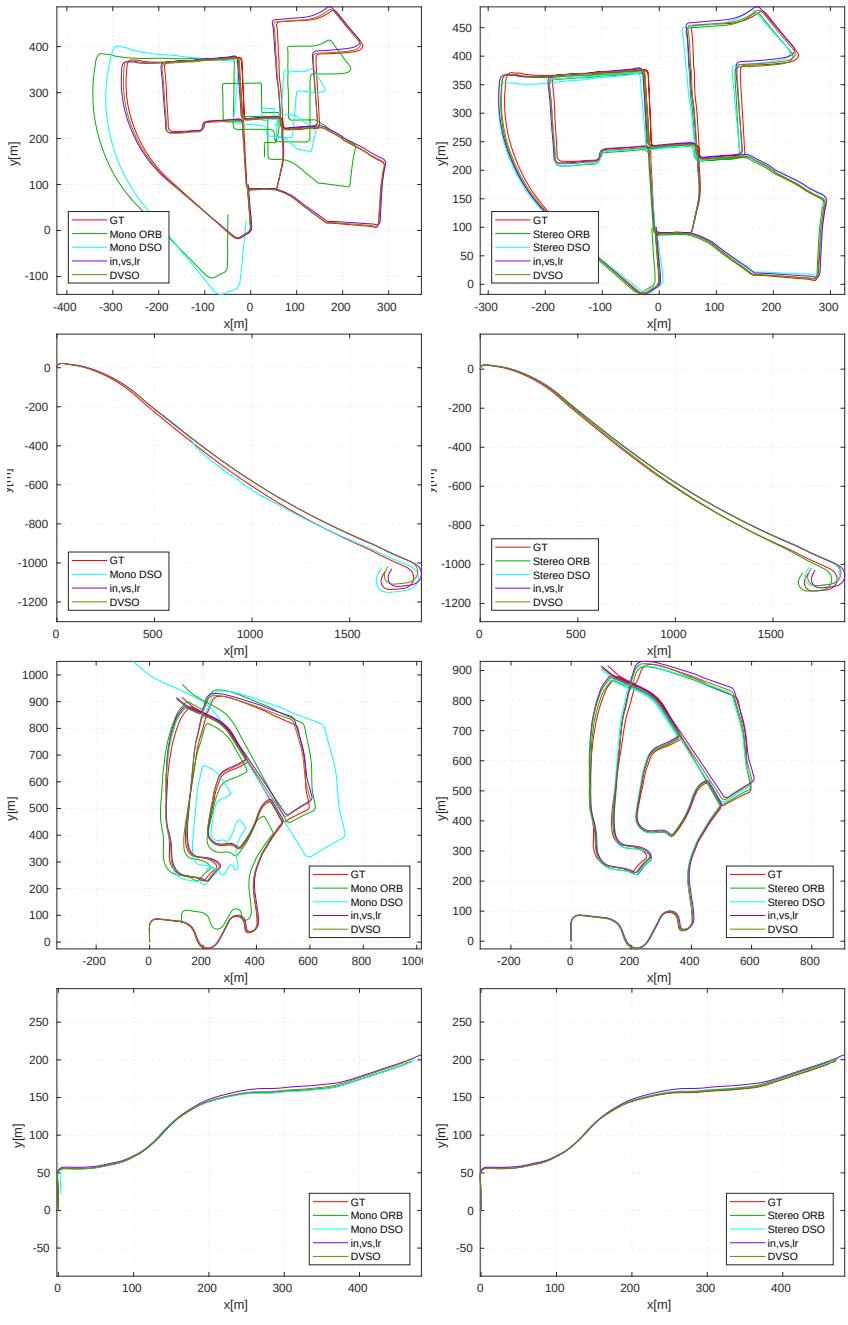


Fig. 5: KITTI seq. 00-03. Note that for sequence 01, the result of Mono ORB is not shown due to its very large scale drift.

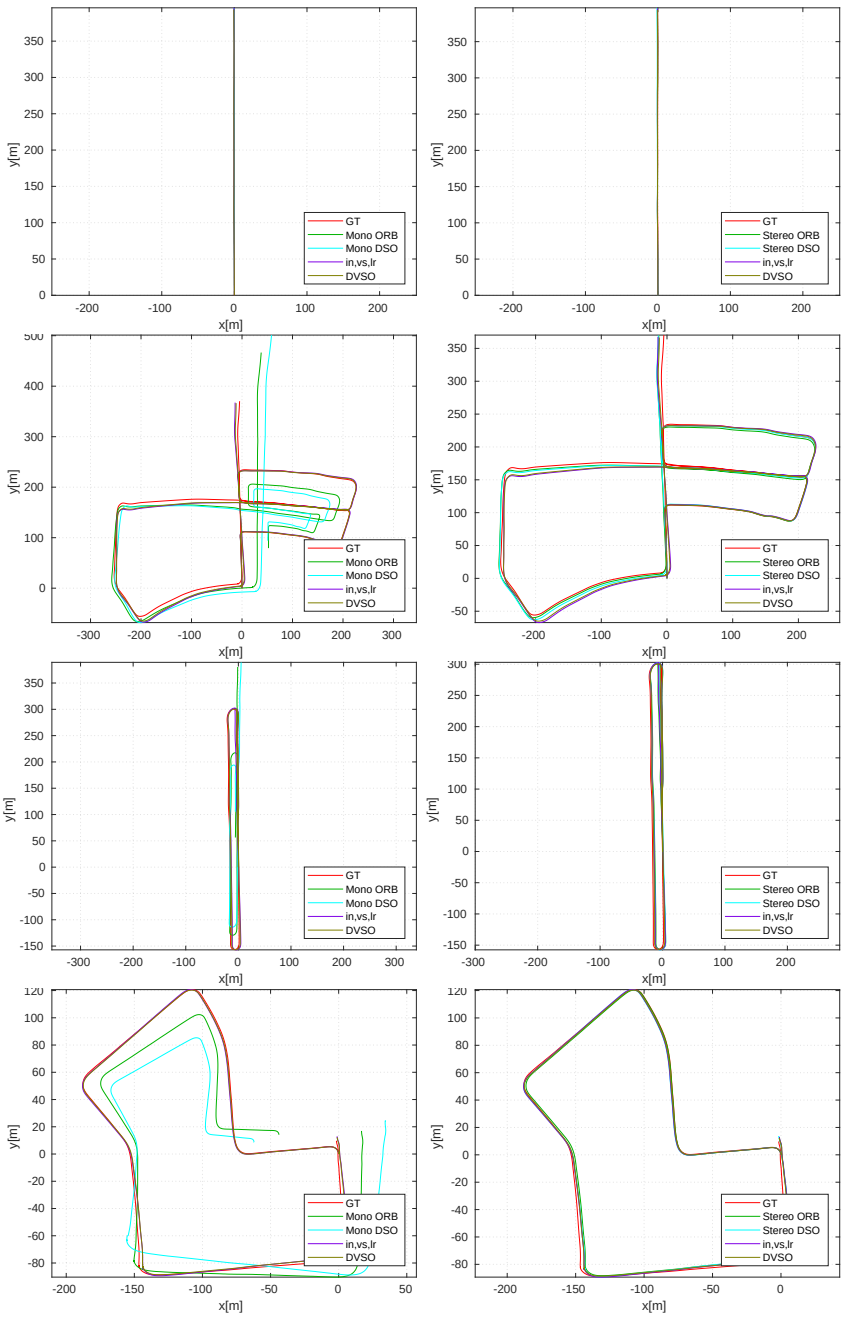


Fig. 6: KITTI seq. 04-07.

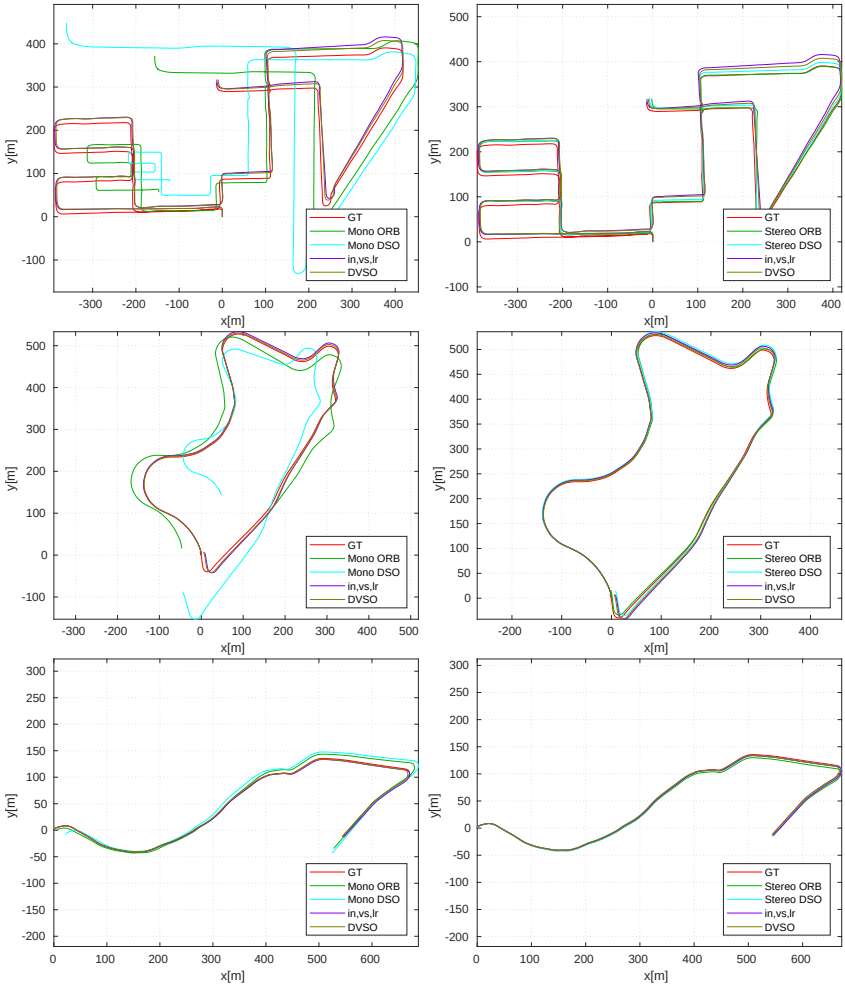


Fig. 7: KITTI seq. 08-10.

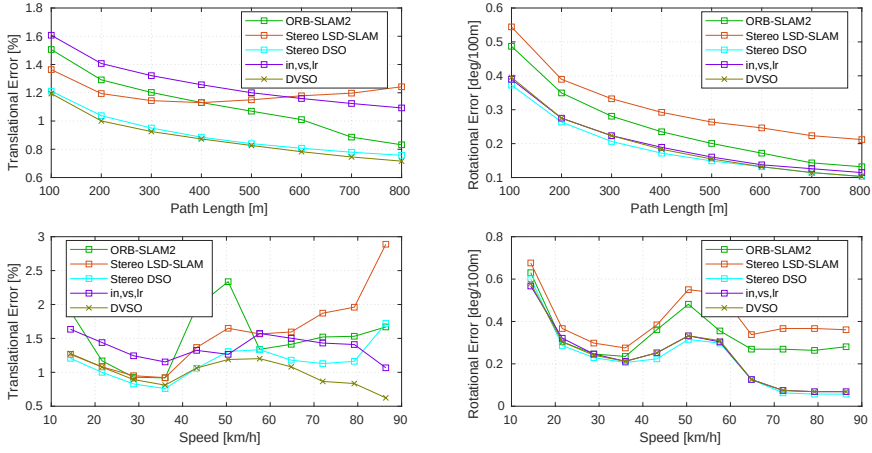


Fig. 8: Evaluation results on the test set of KITTI Odometry Benchmark. The top row shows translational and rotational errors with respect to driving intervals and the bottom row shows the translational and rotational errors with respect to driving speed.

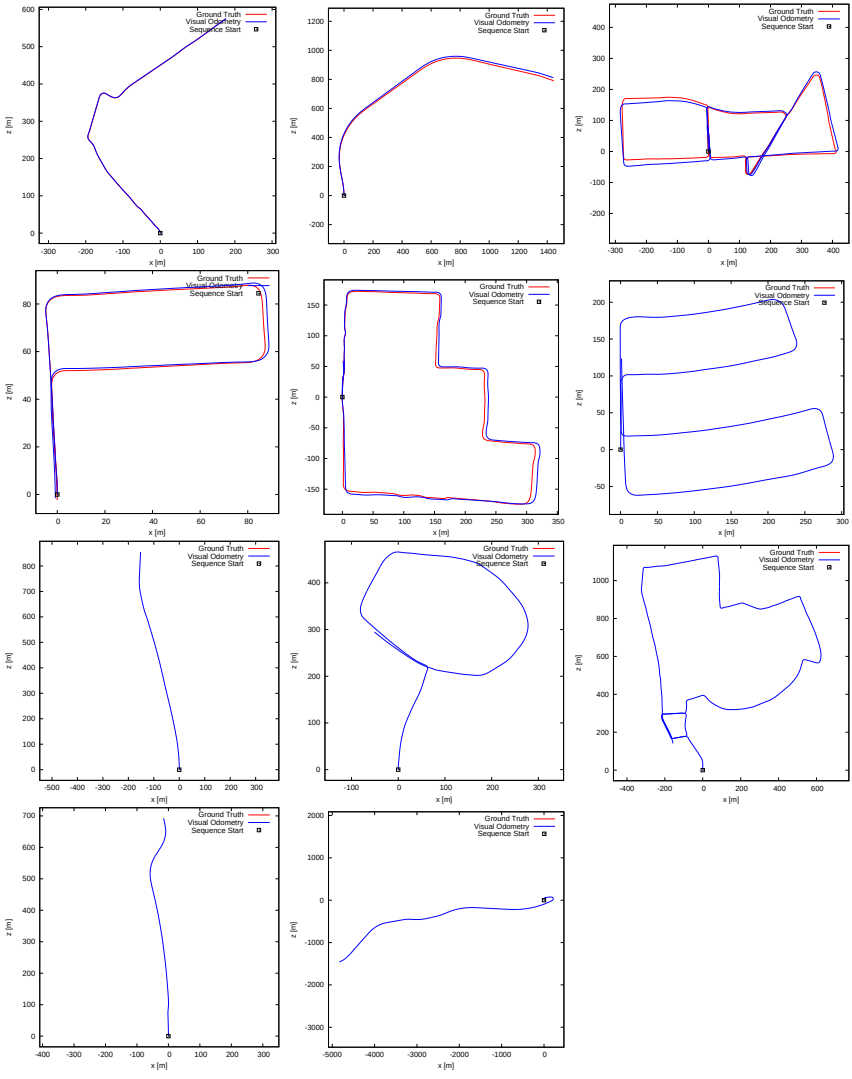


Fig. 9: KITTI seq. 11-21. Note that for seq. 11-15, we downloaded the evaluation results from the KITTI website. For seq. 16-21, one has no access to the ground-truth trajectories.

References

1. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. (2014) 2366–2374
2. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2012)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
4. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. arXiv preprint arXiv:1609.03677 (2016)
5. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE transactions on pattern analysis and machine intelligence (2017)
6. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics **33**(5) (2017) 1255–1262
7. Engel, J., Stückler, J., Cremers, D.: Large-scale direct slam with stereo cameras. In: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, IEEE (2015) 1935–1942
8. Wang, R., Schwörer, M., Cremers, D.: Stereo dso: Large-scale direct sparse visual odometry with stereo cameras. In: International Conference on Computer Vision (ICCV), Venice, Italy (October 2017)