Supplementary Material
Stereo DSO:
Large-Scale Direct Sparse Visual Odometry with Stereo Cameras

Rui Wang*, Martin Schwörer*, Daniel Cremers
Technical University of Munich
{wangr, schwoere, cremers}@in.tum.de

Abstract

*In this **supplementary document**, we first show how weighting the constraints from static stereo differently influences the tracking accuracy. Next, we provide the full trajectory estimations on the training set of KITTI with comparisons to state-of-the-art monocular VO methods. Afterwards we show more results on the Cityscapes Frankfurt sequence, which qualitatively demonstrates the tracking accuracy of our method. We also provide a **supplementary video** to show the performance of our method on the selected datasets, as well as the quality of the delivered 3D reconstructions.*

1. Effect of Stereo Coupling Factor

The estimated trajectories on KITTI Seq. 06 using coupling factors (λ) 0-3 are shown in Fig 1.

2. Full Results on KITTI

Fig 3 shows our trajectory estimates for all training sequences of KITTI (left) and their comparisons to the ground truth (right). To show the improvements over monocular methods, the results of the monocular ORB-SLAM (VO only) and monocular DSO are shown in the middle. The trajectory estimates of the monocular methods are aligned to the ground truth using a similarity transformation (7DoF), while the results of our method are aligned using a rigid-body transformation (6DoF). Obviously, scale drift is the main problem of the monocular methods, which can be resolved by using stereo cameras. In addition, monocular DSO seems to have larger scale drift than monocular ORB-SLAM. We believe this results from the sensitivity of direct methods to the low frame rate, large optical flow, as well as other unmodeled effects in the image domain, such as non-lambertian reflectance and illumination changes that have

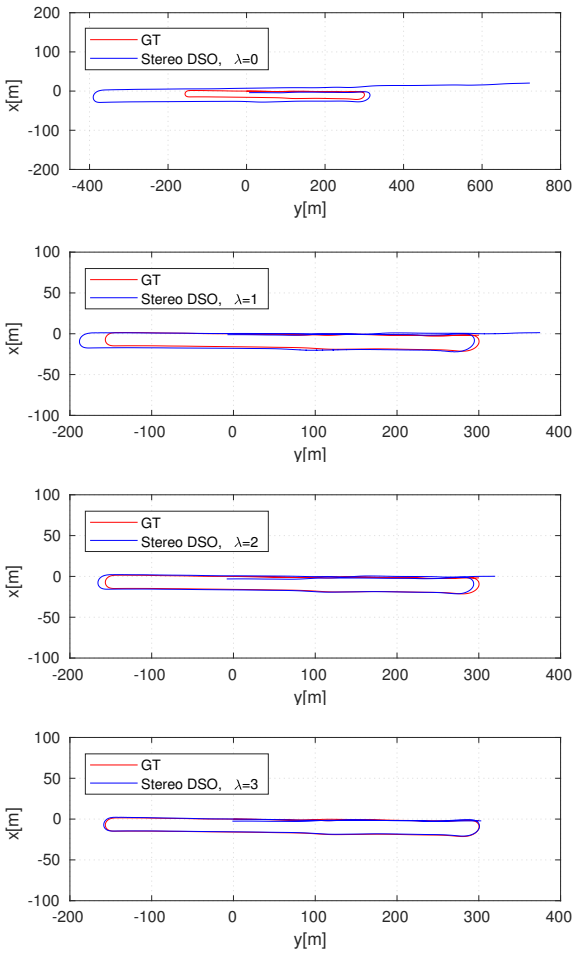


Figure 1: Trajectories on KITTI Seq. 06 using coupling factors (λ) 0-3. Increasing the weighting of the static stereo constraints significantly reduces the translational drift.

*These authors contributed equally.

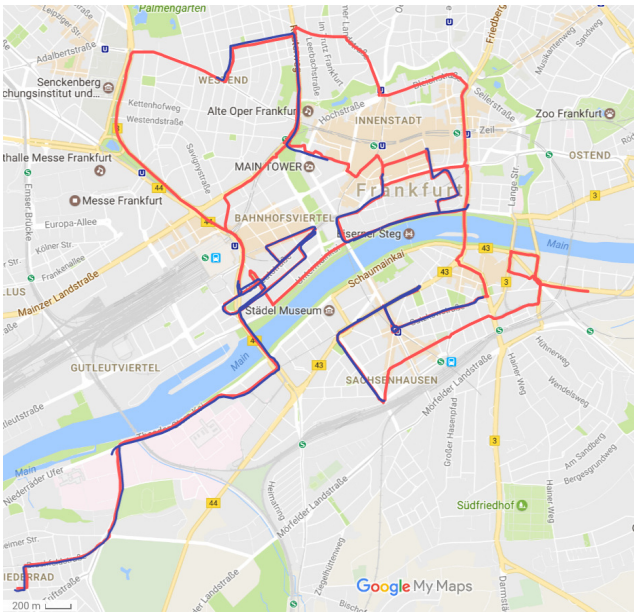
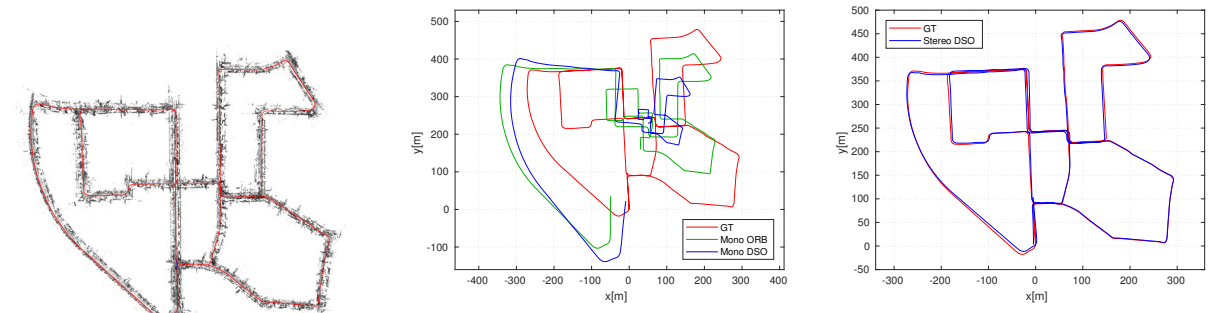


Figure 2: The full Frankfurt trajectory superimposed on the corresponding Google Map scene. The trajectory estimates of the subsections (blue) are aligned to the ground truth trajectory (orange) using a rigid-body transformation (6DoF). Best viewed printed.

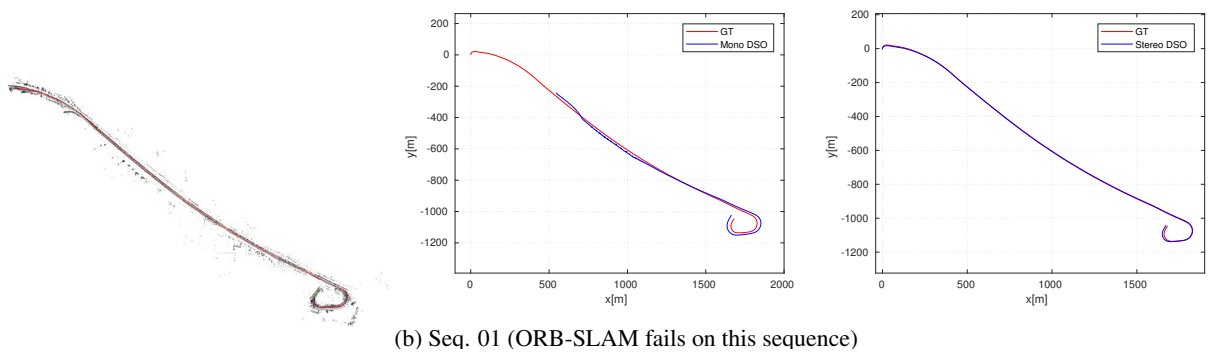
not been corrected sufficiently.

3. More Results on Cityscapes

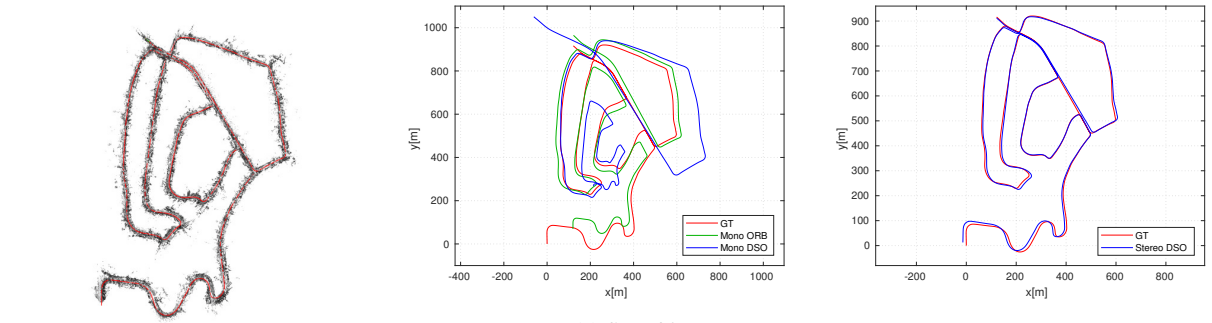
As mentioned in the main paper, without specifically handling moving objects and sudden strong brightness changes, our method is currently not able to run on the entire Frankfurt sequence (around 107,000 frames). Therefore, we divide the full sequence into several smaller sections, each with a length of 5000-6000 frames resulting in a comparable coverage to the KITTI sequences. Exemplary results with ground truth are shown in Fig 4. The plots of a few sections reveal that the ground truth poses calculated from the provided GPS coordinates are not always accurate. In Fig 2 we show the ground truth trajectory of the entire sequence as well as the estimated trajectories aligned to it.



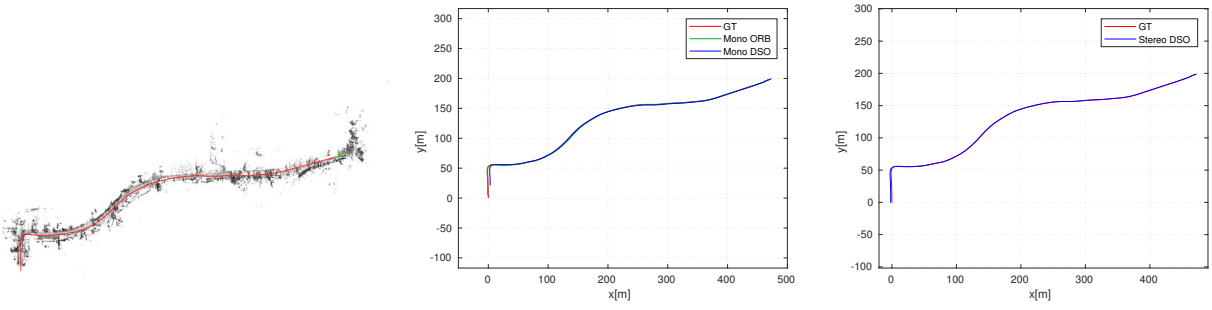
(a) Seq. 00



(b) Seq. 01 (ORB-SLAM fails on this sequence)

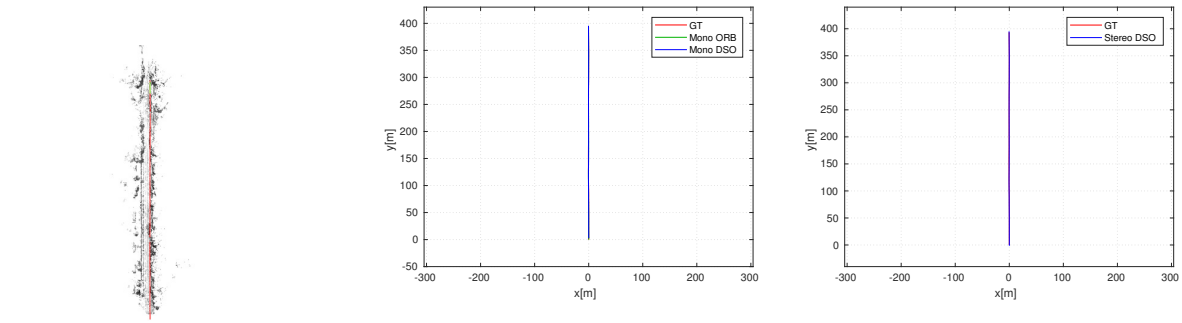


(c) Seq. 02

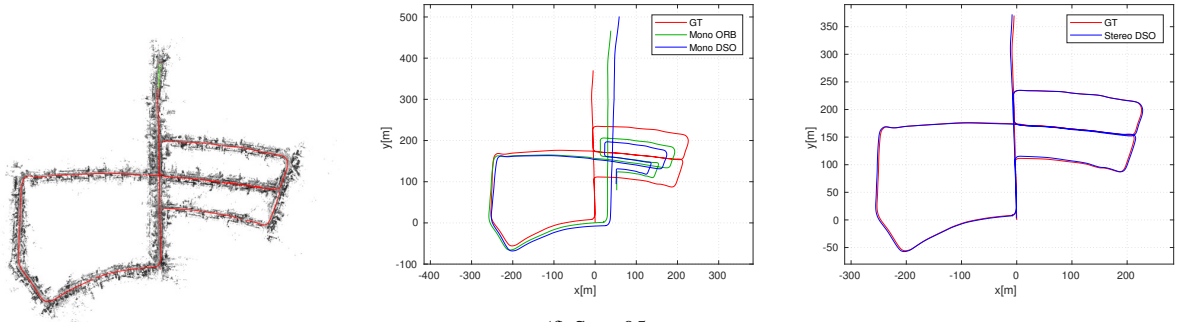


(d) Seq. 03

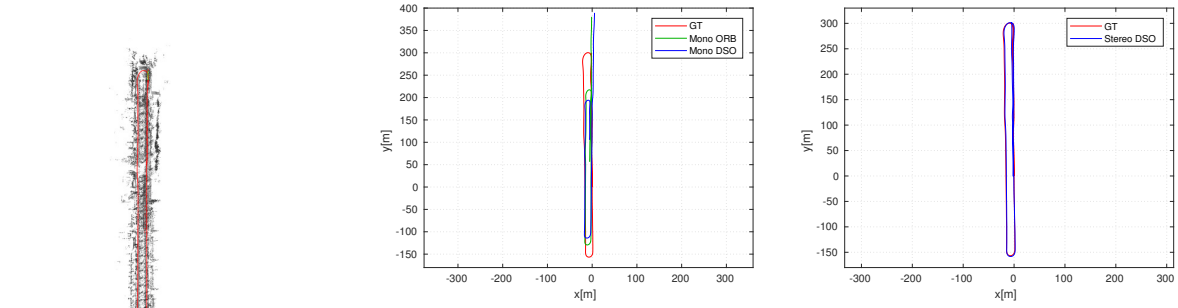
Figure 3: Full results on the KITTI training set. The trajectories estimated by our VO method are shown in the left column. The comparisons to the ground truth are shown in the right column. The results of the two state-of-the-art monocular VO methods, namely ORB-SLAM (VO only) and DSO, are shown in the middle. The trajectories are aligned to the ground truth using similarity transformations (7DoF) and rigid-body transformations (6DoF) for the monocular methods and our stereo method respectively.



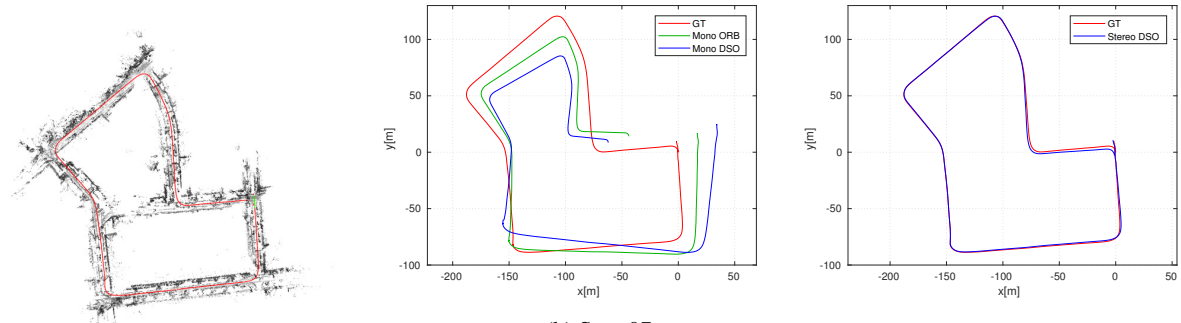
(e) Seq. 04



(f) Seq. 05

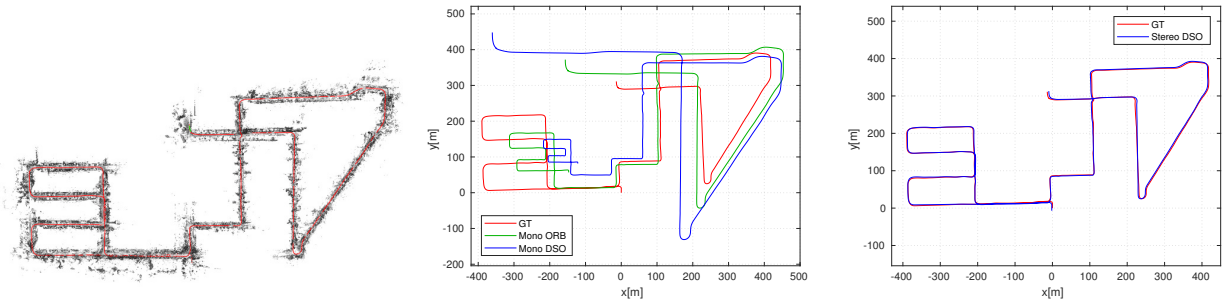


(g) Seq. 06

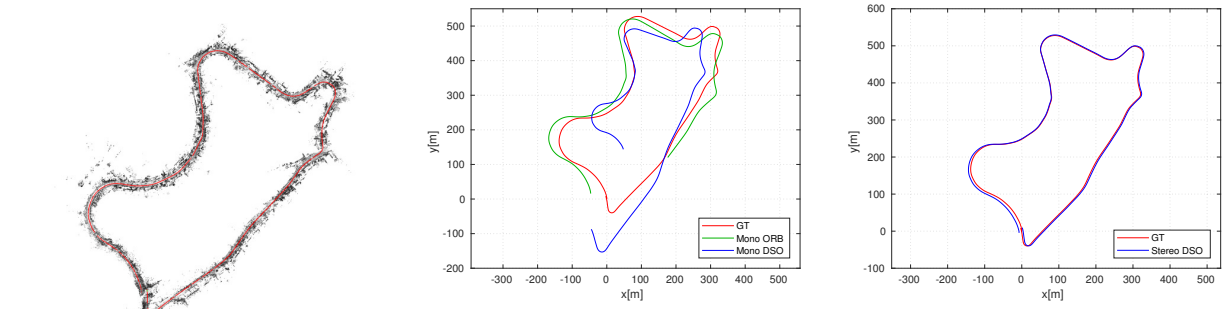


(h) Seq. 07

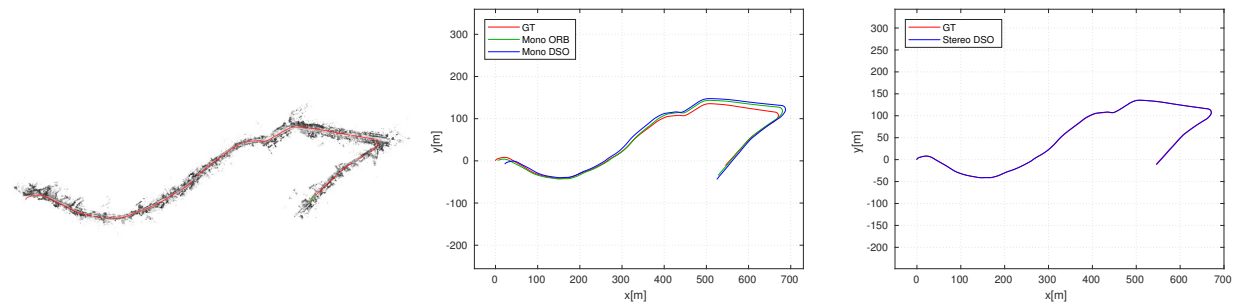
Figure 3: Full results on the KITTI training set (cont.). The trajectories estimated by our VO method are shown in the left column. The comparisons to the ground truth are shown in the right column. The results of the two state-of-the-art monocular VO methods, namely ORB-SLAM (VO only) and DSO, are shown in the middle. The trajectories are aligned to the ground truth using similarity transformations (7DoF) and rigid-body transformations (6DoF) for the monocular methods and our stereo method respectively.



(i) Seq. 08

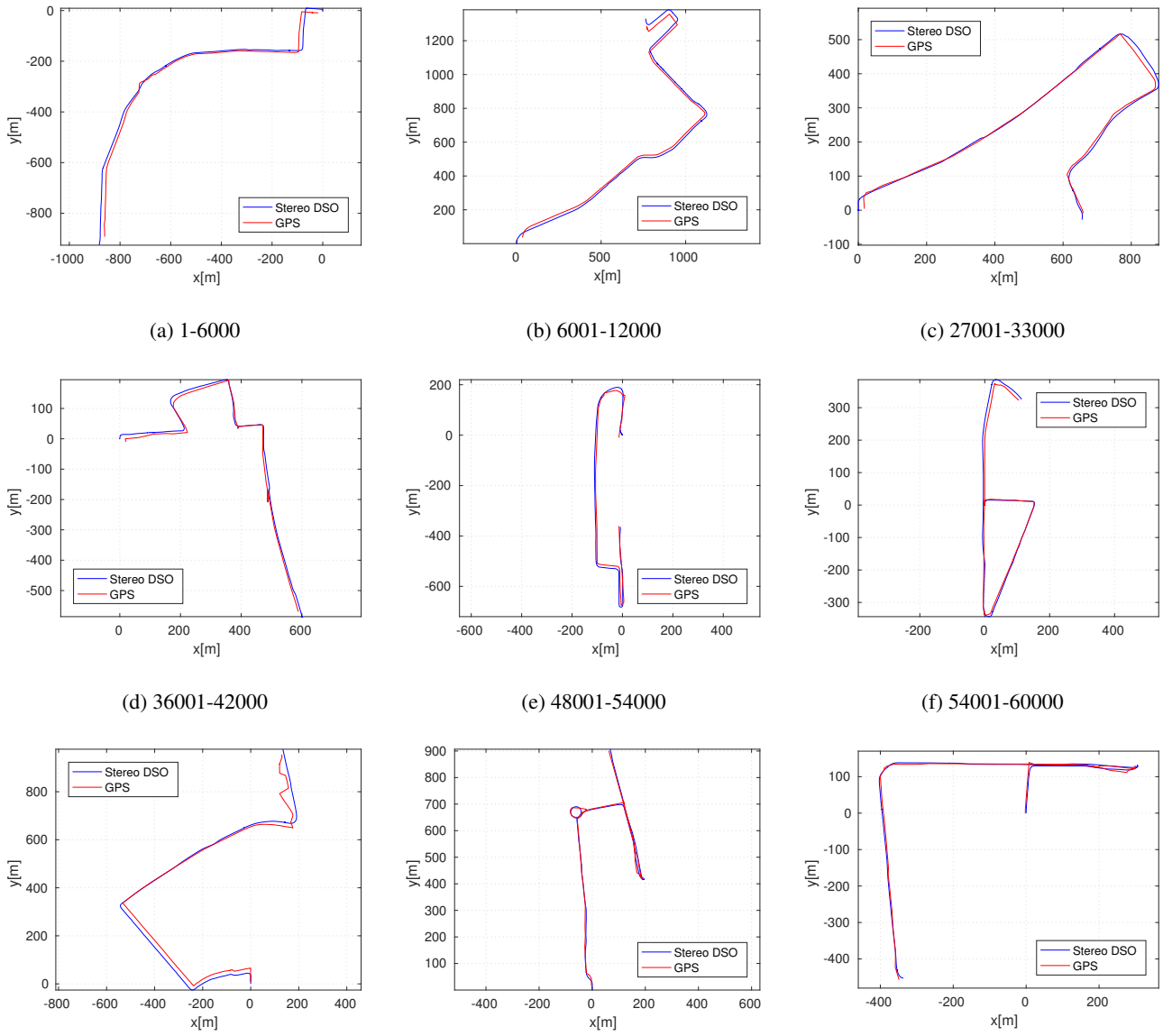


(j) Seq. 09



(k) Seq. 10

Figure 3: Full results on the KITTI training set (cont.). The trajectories estimated by our VO method are shown in the left column. The comparisons to the ground truth are shown in the right column. The results of the two state-of-the-art monocular VO methods, namely ORB-SLAM (VO only) and DSO, are shown in the middle. The trajectories are aligned to the ground truth using similarity transformations (7DoF) and rigid-body transformations (6DoF) for the monocular methods and our stereo method respectively.



(a) 1-6000 (b) 6001-12000 (c) 27001-33000

(d) 36001-42000 (e) 48001-54000 (f) 54001-60000

(g) 69001-55000 (h) 87001-93000 (i) 90001-96000

Figure 4: Estimated camera trajectories on the Cityscapes Frankfurt stereo sequence. The sequences are obtained by dividing the full sequence into several smaller sections with length of 5000 to 6000 frames and coverages comparable to the sequences of KITTI. The sub-captions name the corresponding frame indices in the full sequence. The ground truth poses are calculated from the provided GPS coordinates using the Mercator projection. In some figures, *e.g.* Fig 4(g), the inaccuracies of the GPS are clearly visible.