Depth Estimation for Direct Sparse Visual Odometry

17th Conference on Computer and Robot Vision, Ottawa, Ontario

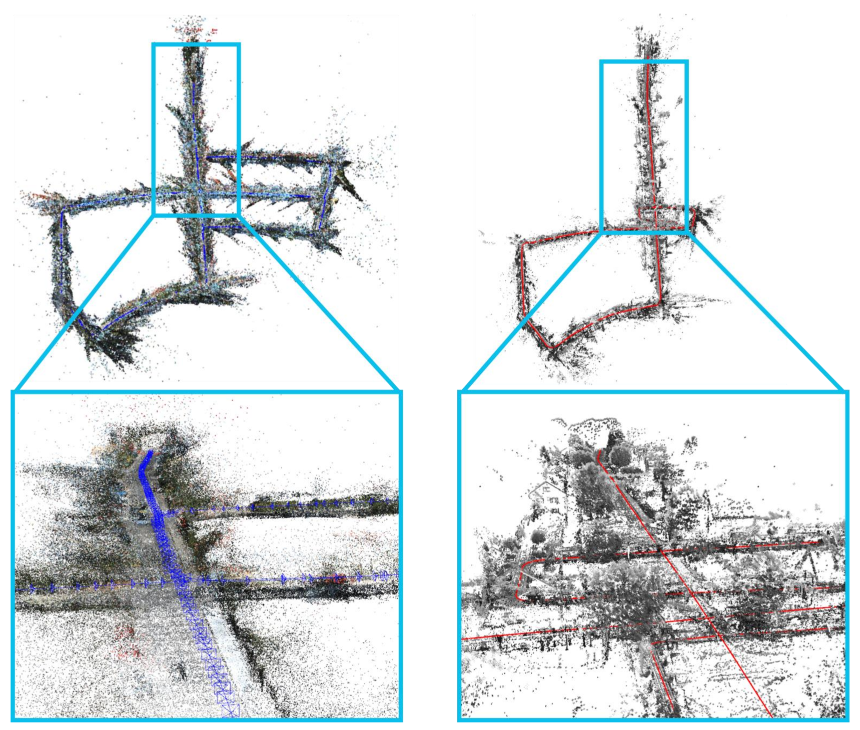
Ran Cheng[1], Christopher Agia[2], David Meger[1], Gregory Dudek[1]

[1]McGill University, [2]University of Toronto

Abstract

We exploit depth prediction as a candidate prior for the coarse initialization, tracking, and marginalization steps of the direct visual odometry system, enabling the second-order optimizer to converge faster into a precise global minimum. In addition, the given depth prior supports large baseline stereo scenarios, maintaining robust pose estimations against challenging motion states such as in-place rotation.

Introduction

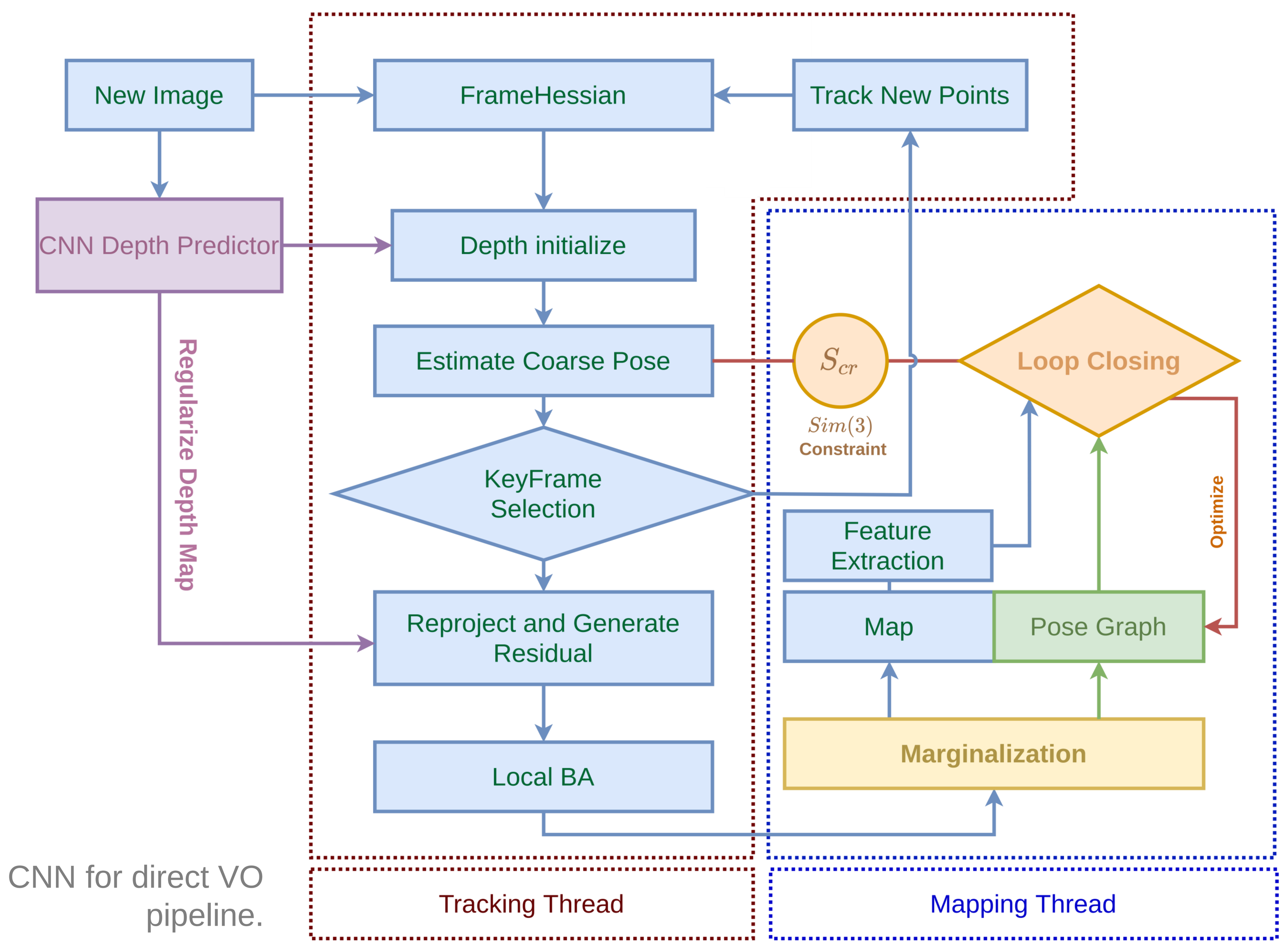


In this work, we present a real-time monocular visual odometry system with an auxiliary deep depth predictor. Since applying the Schur complement for marginalization is equivalent to solving the marginal distribution over camera pose given a depth prior, the joint optimization can thus be broken down into two stages:
a) solve the pose from coarse prior and
b) correct depth from the pose given by stage **a**.

Example mapping results on KITTI seq 05. Left figure is our method, right figure is DSO.

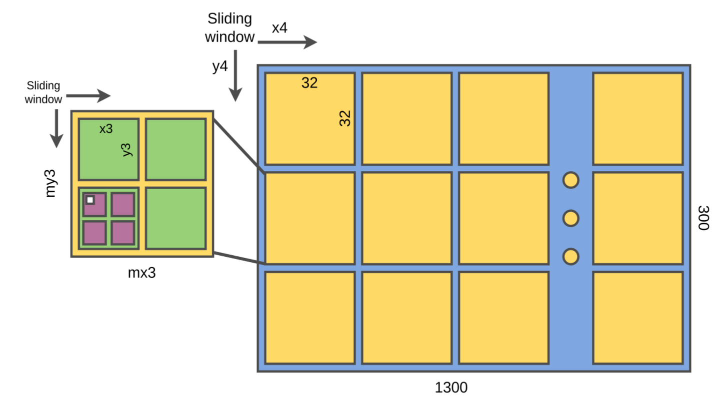
Method

We adopt the keyframe based sliding window approach to estimate the local camera poses and refine the sparse depth maps. Within the local sliding window, the SE(3) keyframe pose and inverse depth are parameterized and composed into a minimization problem. The mutual update of depth and camera poses are optimized through Gauss-Newton method.



Deep CNN for direct VO pipeline.

Depth Estimation in Front End

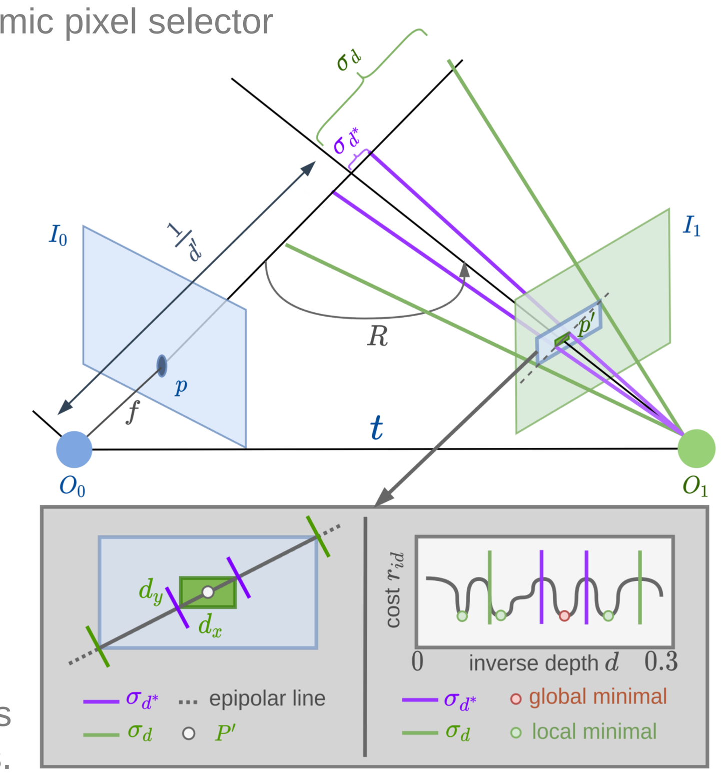


We compute a histogram of gradients within the region defined by a given kernel at a given pixel location. Points are randomly discarded at lower-gradient pixel locations if the region is dense in points.

Dynamic pixel selector

We utilize the pre-trained Monodepth2 [1] model as our depth predictor which helps to refine the pose estimation. By continuously propagating depth into each tracking frame and refining it in the local bundle adjustment, we can easily lock down the inverse depth variance upper bound.

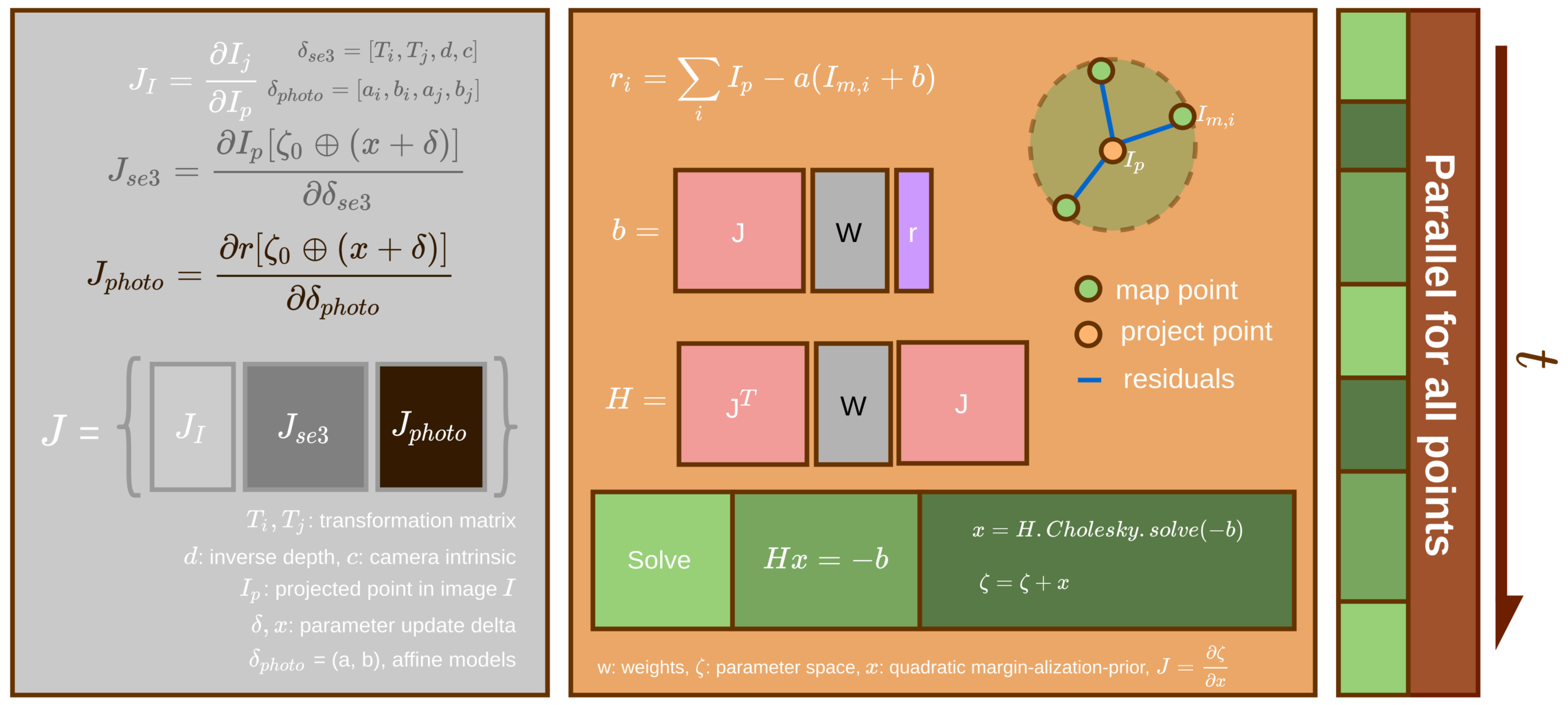
$$J_{id} = I_x d_x + I_y d_y$$

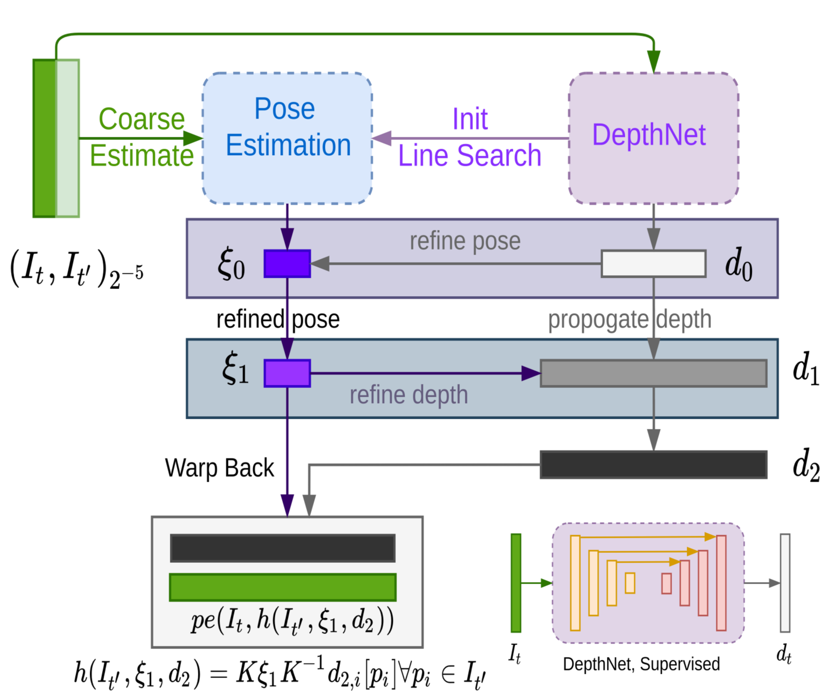


The depth prior narrows the search region and makes our method tenable to large baseline stereo problems.

Depth Estimation in Back-end

the Jacobians of geometrical parameters [2] with respect to two frames' poses are linearly related. The big Hessian matrix can be decomposed into two parts: one pose Hessian which is very small, one point Hessian vector which is a large vector containing all the inverse depth Jacobians in each entry.





The pose Hessian is aggregated through each point Hessian as follows:

$$\hat{H_\xi} = \sum_{i \in \mathcal{P}_g} \{H_\xi - H_{\xi, p_i} H_{p_i}^{-1} H_{p_i, \xi}\}$$

$$\hat{b_\xi} = \sum_{i \in \mathcal{P}_g} \{b_\xi - H_{\xi, p_i} H_{p_i}^{-1} b_{p_i}\}$$

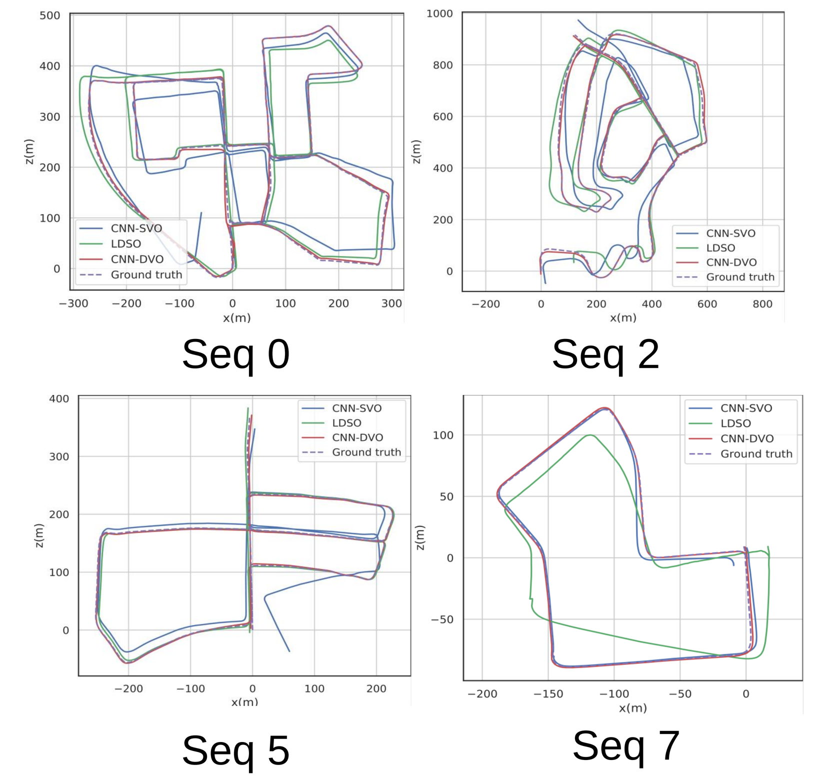
After the linear aggregation, the pose update can be solved from the following equation:

$$\delta_{p_i} = H_{p_i}^{-1}(b_{p_i} - H_{\xi, p_i}^T \xi')$$

Deep CNN for direct VO pipeline.

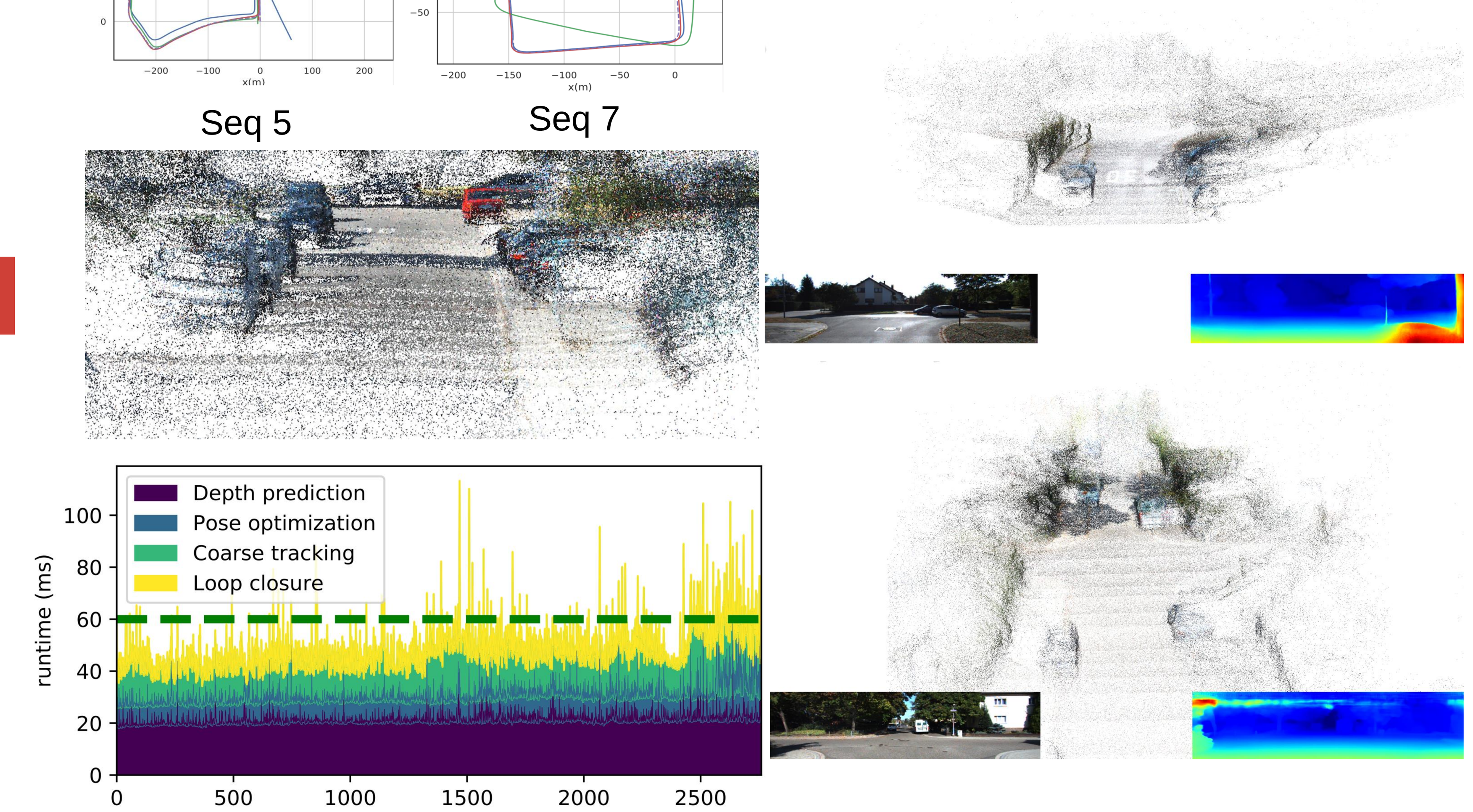
Result

Our experiments show that the inclusion of depth priors helps to progressively reduces scale-drift over long sequences by restricting the reprojection of tracking points in the local bundle adjustment near the global minimum.



Seq	Ours	Ours $w\backslash o$ DP*	Ours $w\backslash o$ LC*	Ours $w\backslash$ PP*	CNN-SVO
00	**4.55**	118.28	13.45	9.34	130.03
01	9.79	9.71	10.48	**6.79**	202.36
02	10.98	125.05	34.92	**5.81**	48.24
03	**1.05**	2.33	2.99	1.25	3.26
04	0.52	0.98	1.12	**0.38**	2.10
05	**2.23**	65.11	4.25	8.97	20.39
06	**3.04**	83.63	13.27	4.89	12.50
07	**0.44**	23.42	2.10	2.35	4.05
08	**4.51**	116.85	6.48	6.49	10.65
09	4.90	68.21	5.80	**4.50**	14.69
10	1.49	13.68	7.68	**1.20**	8.74
mean	**3.96**	57.02	9.14	**5.00**	41.55

$w\backslash o$ DP*: without depth prior ($\mathbf{Sim}(3)$ scale aligned).
$w\backslash o$ LC*: without loop closure for pose optimization.
$w\backslash$ PP*: with pose prior for initialization.



Conclusion

We present a novel deep learning powered direct visual odometry system. To our knowledge, it is the only direct formulation that fully integrates depth prediction with every major system component (initialization, tracking, marginalization) to jointly optimize for all model parameters including inverse depth, camera pose and affine model in real-time (with GPU-based implementation).

Extra constraints given by the depth prior provides strong benefits to our system, as it:
a) fixes the arbitrary initialization scale into constant scale
b) it greatly restrains the scale-drift during tracking
c) it recovers the 3D correspondence of the measurements [3] in loop closure while helping to maintain consistent scale.

References

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

[2] Engel, Jakob, Vladlen Koltun, and Daniel Cremers. "Direct sparse odometry." *IEEE transactions on pattern analysis and machine intelligence* 40.3 (2017): 611-625.

[3] Gao, Xiang, et al. "LDSO: Direct sparse odometry with loop closure." *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018.