Learning Monocular Visual Odometry via Self-Supervised Long-Term Modeling Supplementary Material

Yuliang Zou[1], Pan Ji[2], Quoc-Huy Tran[2], Jia-Bin Huang[1], and Manmohan Chandraker[2,3]

[1]Virginia Tech [2]NEC Labs America [3]UCSD

1 Overview

In this supplementary document, we provide additional experimental results and information to complement the main manuscript. First, we conduct additional ablation experiments to further validate our design choices. Second, we show our results on the KITTI Odometry leaderboard. Third, we show results on the KITTI Odometry training split. Fourth, we show results on the snippet-level pose and single-view depth estimation for completeness. Lastly, we provide the list of sequences we selected from KITTI raw data. We also provide a demo video showing the trajectories of several challenging sequences in the KITTI Odometry dataset. Please refer to the attached file **supp_video.mp4**.

2 Ablation Study

In Table 1, we conduct an ablation study to validate the effectiveness of the incorporated cycle consistency constraint, pose features (from I_0 and I_t), depth features, and the memory buffer in our two-layer ConvLSTM module. As we can see, all the components help improve performance in the first stage of training.

Table 1: **Ablation study on different components** of the second-layer ConvLSTM. The best performance is in **bold** and the second best is underlined.

Method	Seq. 09			Seq. 10		
	RMSE (m)	Rel. trans. (%)	Rel. rot. (deg/m)	RMSE (m)	Rel. trans. (%)	Rel. rot. (deg/m)
Two-layer ConvLSTM (w/o cycle consistency)	20.37	5.02	0.016	16.63	6.88	0.035
Two-layer ConvLSTM (w/o pose features)	14.26	5.64	0.018	14.47	7.52	0.030
Two-layer ConvLSTM (w/o depth features)	11.53	4.54	0.015	14.07	6.54	0.031
Two-layer ConvLSTM (w/o memory buffer)	12.54	5.12	0.014	13.96	7.20	0.026
Two-layer ConvLSTM	**9.77**	**4.23**	**0.013**	**12.68**	**6.02**	**0.023**

In Table 2, we conduct an ablation study to show the performance of different input sequence lengths of the second stage of training. Our results show that the performance gradually improves as we increase the number of input frames during training. When the number of frames reaches the GPU memory limitations (e.g., our default setting, 97-frame), we achieve the best performance. Training the model on a GPU with larger memory could potentially improve the performance further.

Table 2: **Ablation study on different input sequence length** of the second-stage of training. The best performance is in **bold** and the second best is underlined.

Method	Seq. 09			Seq. 10		
	RMSE (m)	Rel. trans. (%)	Rel. rot. (deg/m)	RMSE (m)	Rel. trans. (%)	Rel. rot. (deg/m)
49-frame	12.50	3.83	0.011	12.30	5.99	**0.018**
73-frame	12.42	3.69	**0.010**	12.06	5.89	**0.018**
97-frame (default)	**11.30**	**3.49**	**0.010**	**11.80**	**5.81**	**0.018**

3 Results on KITTI Odometry Test Set

In Table 3, we provide results on the KITTI Odometry leaderboard. It may be observed that the performance of our method is close to Table 5 in the main manuscript. This suggests that using ORB-SLAM2-S as pseudo ground truth is a reasonable choice for evaluation.

In addition to our method, we select two state-of-the-art self-supervised methods (CC [8] and MonoDepth2 [5]) and submit the estimated results to the server as well. Our method compares favorably with these two self-supervised methods. Our method also outperforms the supervised method DeepVO [12] by a large margin.

Table 3: **Results on KITTI Odometry leaderboard.** Note that we use the estimations from ORB-SLAM2-S [7] to align scale globally for the self-supervised methods.

	Method	Rel. trans (%)	Rel. rot. (deg/m)
Geo.	ORB-SLAM2-S [7]	1.70	0.0028
	VISO2-M [4]	11.94	0.0234
	VISO2-M+GP [4,10]	7.46	0.0245
Sup.	DeepVO [12]	24.55	0.0489
Self-Sup.	CC [8]	16.06	0.0320
	MonoDepth2 [5]	12.59	0.0312
	Ours	7.40	0.0142

In Figure 1, we show qualitative results on the remaining 7 sequences (other than those shown in the main manuscript) from the KITTI Odometry test set. Our method aligns best with the reference ORB-SLAM2-S trajectories.

4 Results on KITTI Odometry Training Set

In Table 4, we compare the results on the training set of the KITTI Odometry dataset. Note that all supervised methods are trained on Sequence 00, 02, 08,

Table 4: **Pose evalution** on *training split* of KITTI Odometry dataest [3]. The results of ORB-SLAM2-M methods are the medians of 5 times. '-' means the results are not available from that paper. For DeepV2D [11], SfMLearner [17], GeoNet [16], CC [8], DeepMatchVO [9], and MonoDepth2 [5], we take the pre-trained models and run on the sequences to get the results. The best performance of each block is in **bold**, and the second best is underlined.

	RMSE (m)	Seq. 00	Seq. 01	Seq. 02	Seq. 03	Seq. 04	Seq. 05	Seq. 06	Seq. 07	Seq. 08
Geo.	ORB-SLAM2-M (w/o LC)	54.94	568.63	58.55	**1.41**	2.41	29.32	51.87	16.83	**36.90**
	ORB-SLAM2-M	**9.02**	**529.28**	**17.96**	2.07	**1.56**	**5.20**	**14.07**	**2.88**	37.83
Sup.	DeepV2D [11]	101.08	484.87	121.02	3.62	8.86	35.23	113.31	12.86	55.69
Self-Sup.	SfMLearner [17]	97.81	108.09	152.15	7.47	2.49	48.13	39.56	21.28	32.56
	GeoNet [16]	148.81	168.90	293.46	17.58	7.26	86.94	17.69	13.88	138.00
	CC [8]	68.31	50.41	59.19	8.89	2.25	22.49	13.02	11.31	49.29
	DeepMatchVO [9]	51.34	85.96	127.99	11.03	3.09	27.59	20.98	16.71	38.71
	MonoDepth2 [5]	82.05	**30.81**	86.64	2.40	**2.00**	21.49	**5.16**	10.42	51.83
	Ours	**13.13**	41.38	**12.61**	**1.61**	2.22	**8.24**	9.16	**9.92**	**13.98**

	Rel. trans (%)	Seq. 00	Seq. 01	Seq. 02	Seq. 03	Seq. 04	Seq. 05	Seq. 06	Seq. 07	Seq. 08
Geo.	ORB-SLAM2-M (w/o LC)	14.11	131.75	12.70	**1.21**	2.40	9.12	18.50	10.34	**9.72**
	ORB-SLAM2-M	**3.23**	**125.63**	**3.69**	1.73	**1.97**	**2.31**	**5.92**	**2.15**	11.68
Sup.	DeepVO [12]	-	-	-	8.49	7.19	2.62	5.42	3.91	-
	ESP-VO [13]	-	-	-	6.72	6.33	3.35	7.24	3.52	-
	GFS-VO [14]	-	-	-	5.44	**2.91**	3.27	8.50	3.37	-
	GFS-VO-RNN [14]	-	-	-	6.36	5.95	5.85	14.58	5.88	-
	BeyondTracking [15]	-	-	-	**3.32**	2.96	**2.59**	**4.93**	**3.07**	-
	DeepV2D [11]	**12.38**	**56.26**	**7.79**	4.07	8.22	6.35	16.67	4.96	**6.63**
Self-Sup.	SfMLearner [17]	19.27	21.71	18.99	9.73	3.17	10.02	11.00	11.68	8.67
	GeoNet [16]	33.63	22.96	54.00	19.41	10.81	22.68	9.90	9.82	22.26
	CC [8]	10.42	15.64	8.08	8.49	2.90	5.70	4.38	5.91	7.16
	DeepMatchVO [9]	5.31	29.57	15.94	9.67	4.15	7.42	5.69	7.62	9.43
	MonoDepth2 [5]	7.64	**10.06**	8.34	5.30	3.20	4.66	**2.48**	**4.58**	7.32
	Ours	**2.60**	13.27	**2.49**	**1.59**	**2.52**	**2.63**	2.64	6.43	**3.61**

	Rel. rot (deg/m)	Seq. 00	Seq. 01	Seq. 02	Seq. 03	Seq. 04	Seq. 05	Seq. 06	Seq. 07	Seq. 08
Geo.	ORB-SLAM2-M (w/o LC)	**0.003**	**0.010**	**0.003**	**0.002**	**0.002**	**0.002**	0.003	**0.003**	**0.003**
	ORB-SLAM2-M	**0.003**	0.012	0.004	**0.002**	**0.002**	0.003	**0.002**	0.005	**0.003**
Sup.	DeepVO [12]	-	-	-	0.069	0.070	0.036	0.058	0.046	-
	ESP-VO [13]	-	-	-	0.065	0.061	0.049	0.073	0.050	-
	GFS-VO [14]	-	-	-	0.033	**0.013**	0.016	0.027	0.022	-
	GFS-VO-RNN [14]	-	-	-	0.036	0.024	0.025	0.050	0.026	-
	BeyondTracking [15]	-	-	-	**0.021**	0.018	**0.012**	**0.019**	**0.018**	-
	DeepV2D [11]	**0.051**	**0.051**	**0.030**	**0.021**	0.034	0.027	0.073	0.030	**0.031**
Self-Sup.	SfMLearner [17]	0.057	0.026	0.033	0.035	0.033	0.036	0.038	0.059	0.026
	GeoNet [16]	0.057	0.041	0.061	0.098	0.070	0.077	0.043	0.059	0.078
	CC [8]	0.035	0.011	0.016	0.041	0.012	0.022	0.008	0.031	0.023
	DeepMatchVO [9]	0.013	0.013	0.024	0.046	0.020	0.017	0.022	0.037	0.012
	MonoDepth2 [5]	0.021	0.010	0.015	0.014	0.008	0.017	**0.004**	0.026	0.024
	Ours	**0.005**	**0.003**	**0.003**	**0.006**	**0.005**	**0.005**	0.007	**0.021**	**0.003**

09 of the KITTI Odometry dataset [3], except DeepV2D [11], which is trained on the Eigen split of KITTI raw dataset [2]. Comparing to other self-supervised approaches, our method achieves smaller errors on the training set, indicating that the proposed system can effectively learn to model the camera pose trajectory during training time. Our method also compares favorably against the geometric-based method ORB-SLAM2.

In Figure 2, we show the qualitative results of our method on Seq. 00-08 on the KITTI Odometry dataset.

5 Snippet-level Pose Results and Depth Results

For completeness, we provide the pose estimation results when evaluating on 5-frame snippets in Table 5 and the single-view depth estimation results in Table 6. Note that the depth network is fixed during the second stage of training, so for the depth evaluation, we only train our model for the first stage on the Eigen split of the KITTI raw dataset. As we can see in Table 5, although CC [8] and DeepMatchVO [9] achieve good results on the snippet-level, their results on the video-level are no longer the state-of-the-art. This indicates that evaluating camera pose estimation performance on the snippet-level could be inaccurate, and thus we need to evaluate the whole trajectory to reflect the holistic performance. In Table 6, we also observe that our method slightly outperforms the current self-supervised state-of-the-art MonoDepth2 [5], which indicates that a better pose estimation module could lead to a better depth estimation performance.

Table 5: **5-frame snippet-level results** on KITTI Odometry datest [3].

	Seq. 09	Seq. 10
ORB-SLAM (full)	0.014±0.008	0.012±0.011
SfMLearner [17]	0.021±0.017	0.020±0.015
vid2depth [6]	0.013±0.010	0.012±0.011
GeoNet [16]	0.012±0.007	0.012±0.009
DF-Net [18]	0.017±0.007	0.015±0.009
CC [8]	0.012±0.007	0.012±0.008
DeepMatchVO [9]	0.009±0.005	0.008±0.007
MonoDepth2 [5]	0.017±0.008	0.015±0.010
Ours	0.015±0.006	0.015±0.009

6 Additional KITTI Sequences

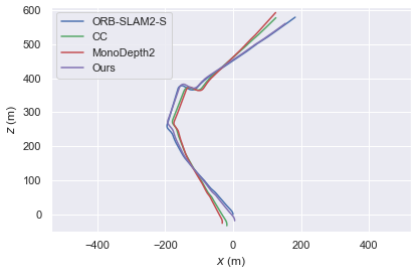
As mentioned in the main manuscript, we selected 18 sequences from KITTI raw data to further evaluate the methods, which have no overlaps with either KITTI Odometry split or Eigen split. We list the sequence names in Table 7.

Table 6: **Single-view depth estimation results** on *Eigen test split* of KITTI raw dataset [2].

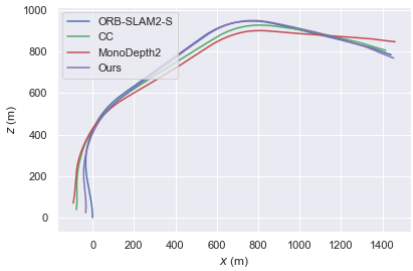
	Error metric ↓				Accuracy metric ↑		
Method	Abs Rel	Sq Rel	RMSE	log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SfMLearner [17]	0.208	1.768	6.856	0.283	0.678	0.885	0.957
vid2depth [6]	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GeoNet [16]	0.155	1.296	5.857	0.233	0.793	0.931	0.973
DF-Net [18]	0.150	1.124	5.507	0.223	0.806	0.933	0.973
CC [8]	0.140	1.070	5.326	0.217	0.826	0.941	0.975
DeepMatchVO [9]	0.156	1.309	5.73	0.236	0.797	0.929	0.969
MonoDepth2 [5]	0.115	0.903	4.863	0.193	0.877	0.959	0.981
SC-SfMLearner [1]	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Ours	0.115	0.871	4.778	0.191	0.874	0.961	0.982

Table 7: **Names of 18 additional KITTI sequences**.

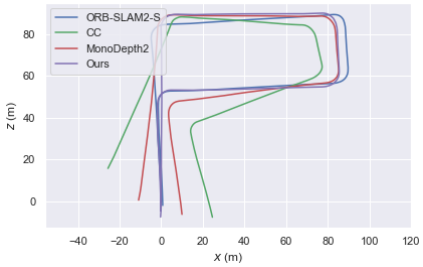
Sequence names
2011_09_26_drive_0036
2011_09_26_drive_0086
2011_09_26_drive_0101
2011_09_26_drive_0117
2011_09_29_drive_0071
2011_10_03_drive_0047
2011_09_26_drive_0059
2011_09_26_drive_0027
2011_09_26_drive_0009
2011_09_26_drive_0013
2011_09_26_drive_0029
2011_09_26_drive_0064
2011_09_26_drive_0084
2011_09_26_drive_0096
2011_09_26_drive_0106
2011_09_26_drive_0056
2011_09_26_drive_0023
2011_09_26_drive_0093



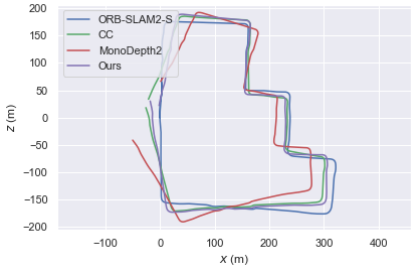
(a) Seq. 11



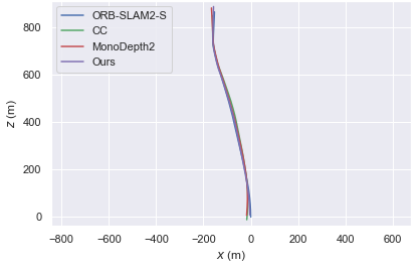
(b) Seq. 12



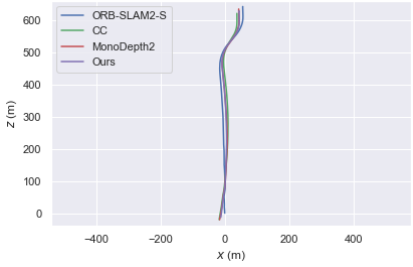
(a) Seq. 14



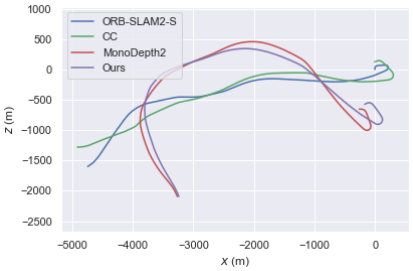
(b) Seq. 15



(a) Seq. 17

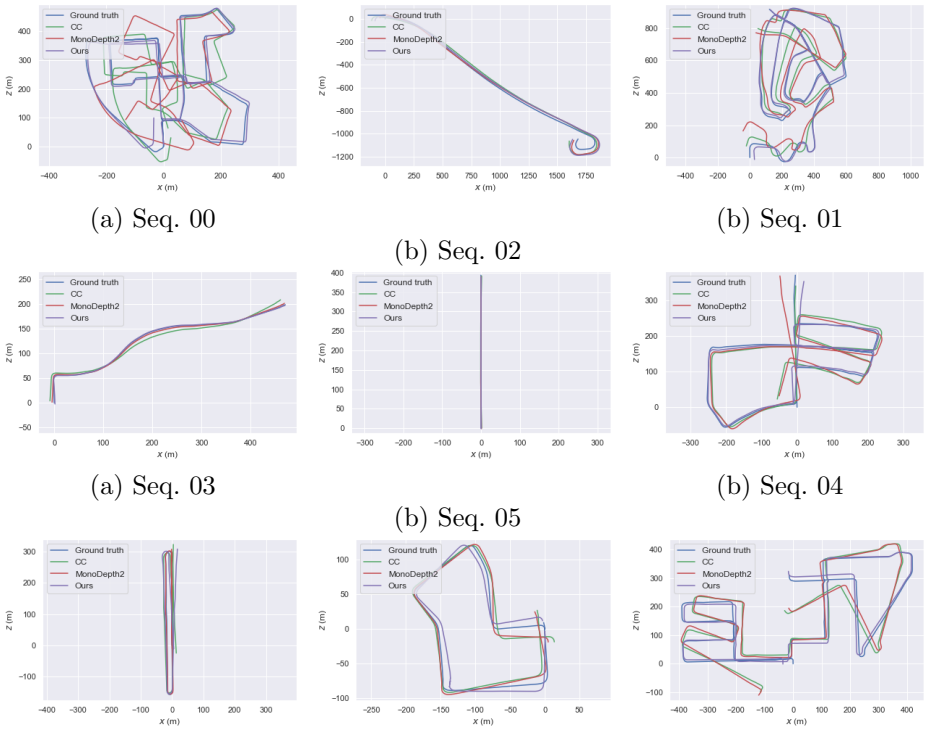


(b) Seq. 20



(b) Seq. 21

Fig. 1: **Visual comparison on the KITTI Odometry test set.** We show the trajectories of ORB-SLAM2-S, CC [8], MonoDepth2 [5] and our method. Our method aligns best with the reference ORB-SLAM2-S trajectories.



(a) Seq. 00 (b) Seq. 02 (b) Seq. 01

(a) Seq. 03 (b) Seq. 05 (b) Seq. 04

(a) Seq. 06 (b) Seq. 08 (b) Seq. 07

Fig. 2: **Visual comparison on the KITTI Odometry training set.** We show the trajectories of ORB-SLAM2-M, CC [8], MonoDepth2 [5] and our method.

References

1. Bian, J.W., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.M., Reid, I.: Unsupervised scale-consistent depth and ego-motion learning from monocular video. In: NeurIPS (2019) 5
2. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. IJRR **32**(11), 1231–1237 (2013) 4, 5
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012) 3, 4
4. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3d reconstruction in real-time. In: 2011 IEEE Intelligent Vehicles Symposium (IV) (2011) 2
5. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.: Digging into self-supervised monocular depth estimation. In: ICCV (2019) 2, 3, 4, 5, 6, 7
6. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: CVPR (2018) 4, 5
7. 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), 1255–1262 (2017) 2
8. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: CVPR (2019) 2, 3, 4, 5, 6, 7
9. Shen, T., Luo, Z., Zhou, L., Deng, H., Zhang, R., Fang, T., Quan, L.: Beyond photometric loss for self-supervised ego-motion estimation. In: ICRA (2019) 3, 4, 5
10. Song, S., Chandraker, M.: Robust scale estimation in real-time monocular sfm for autonomous driving. In: CVPR (2014) 2
11. Teed, Z., Deng, J.: Deepv2d: Video to depth with differentiable structure from motion. In: ICLR (2020) 3, 4
12. Wang, S., Clark, R., Wen, H., Trigoni, N.: Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: ICRA (2017) 2, 3
13. Wang, S., Clark, R., Wen, H., Trigoni, N.: End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. IJRR **37**(4-5), 513–542 (2018) 3
14. Xue, F., Wang, Q., Wang, X., Dong, W., Wang, J., Zha, H.: Guided feature selection for deep visual odometry. In: ACCV (2018) 3
15. Xue, F., Wang, X., Li, S., Wang, Q., Wang, J., Zha, H.: Beyond tracking: Selecting memory and refining poses for deep visual odometry. In: CVPR (2019) 3
16. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: CVPR (2018) 3, 4, 5
17. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017) 3, 4, 5
18. Zou, Y., Luo, Z., Huang, J.B.: DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In: ECCV (2018) 4, 5