Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video

Jia-Wang Bian[12], Zhichao Li[3], Naiyan Wang[3], Huangying Zhan[12], Chunhua Shen[12], Ming-Ming Cheng[4], Ian Reid[12]
[1]University of Adelaide, [2]Australian Centre for Robotic Vision, [3]TuSimple, [4]Nankai University

Problem

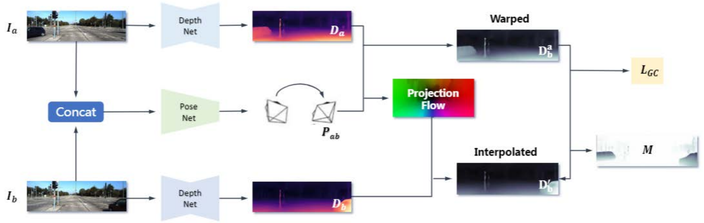
SfMLearner [1] provides a great baseline for unsupervised learning of depth and ego-motion using monocular video. However,

- It produces scale-inconsistent predictions
- The performance is limited due to dynamics and occlusions

Contribution

- Geometry-Consistency loss for scale-consistency
- Self-discovered Mask for handling dynamics and occlusions

Learning Framework



$$L = \alpha L_p^M + \beta L_s + \gamma L_{GC},$$

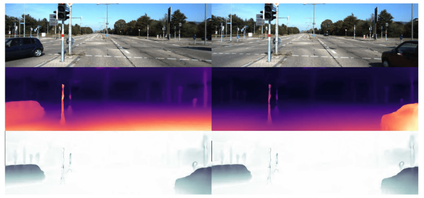
Proposed GC and Mask

$$D_{\text{diff}}(p) = \frac{|D_b^a(p) - D_b'(p)|}{D_b^a(p) + D_b'(p)}$$

$$L_{GC} = \frac{1}{|V|} \sum_{p \in V} D_{\text{diff}}(p),$$

$$M = 1 - D_{\text{diff}},$$

Visualization of Depth and Mask



Depth Results on KITTI

Methods	AbsRel ↓	Acc (<1.25) ↑
SfMLearner [1]	0.208 (0.198)	0.678 (0.718)
CC [2]	0.140 (0.139)	0.826 (0.827)
Ours	**0.137 (0.128)**	**0.830 (0.846)**

- () indicates pretraining on Cityscapes dataset
- SfMLearner [1] is our baseline
- CC[2] is previous SOTA method that jointly learns depth, ego-motion, optical flow, and mask segmentation
- CC [1] needs 7 days for training, while 32 hours for our method

Depth Results with different network and resolution

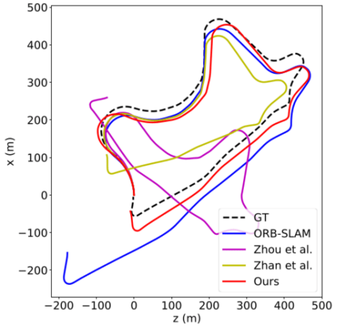
Methods	Resolutions	Error ↓ AbsRel	SqRel	RMS	RMSlog	Accuracy ↑ < 1.25	$< 1.25^2$	$< 1.25^3$
DispNet	416×128	0.151	1.154	5.716	0.232	0.798	0.930	0.972
DispResNet	416×128	0.149	1.137	5.771	0.230	0.799	0.932	0.973
DispNet	832×256	0.146	1.197	5.578	0.223	0.814	0.940	**0.975**
DispResNet	832×256	**0.137**	**1.089**	**5.439**	**0.217**	**0.830**	**0.942**	**0.975**

Inference Time (per image or pair)

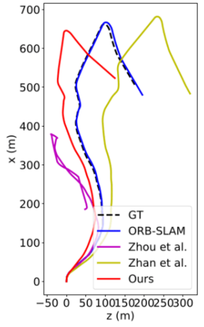
	DispNet	DispResNet	PoseNet
128×416	4.9 ms	9.6 ms	0.6 ms
256×832	9.2 ms	15.5 ms	1.0 ms

Visual Odometry Results

Methods	Seq. 09 t_{err} (%)	Seq. 09 r_{err} ($^\circ$/100m)	Seq. 10 t_{err} (%)	Seq. 10 r_{err} ($^\circ$/100m)
ORB-SLAM [10]	15.30	0.26	3.68	0.48
Zhou et al. [5]	17.84	6.78	37.91	17.78
Zhan et al. [16]	11.93	3.91	12.45	**3.46**
Ours (K)	11.2	3.35	**10.1**	4.96
Ours (CS+K)	**8.24**	**2.19**	10.7	4.58



(a) sequence 09



(b) sequence 10

Reference

- [1] Zhou et al. Unsupervised learning of depth and ego-motion from video. In CVPR, 2017.
- [2] Ranjan et al. Competitive Collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In CVPR, 2019.

Code and Paper

- Scan the QR code.
- Google “SC-SfMLearner”