无监督且尺度一致的深度估计与视觉SLAM

讲者：边佳旺
单位：阿德莱德大学

分享大岗

1. 单目无监督深度估计原理
2. 输出尺度不一致问题
3. 我们的解决方案
4. 用输出尺度一致的深度做视觉SLAM
5. 三维重构Demo

1. 单目无监督深度估计原理

根据训练数据分为三类：

1. Stereo Pair

   (Grag et al, ECCV 2016) Unsupervised CNN for Single View Depth Estimation: Geometry to Rescue

2. Monocular Video

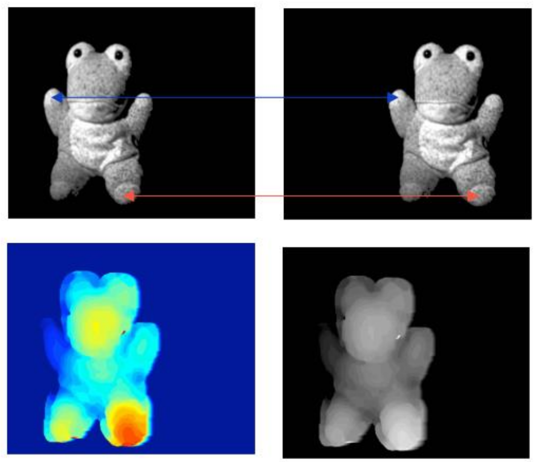
   (SfMLearner, CVPR 2017) Unsupervised Learning of Depth and Ego-Motion from Video

3. Stereo Video

   (Depth-VO-Feat, CVPR 18) Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction

1. 单目无监督深度估计原理

- Recover depth from two views

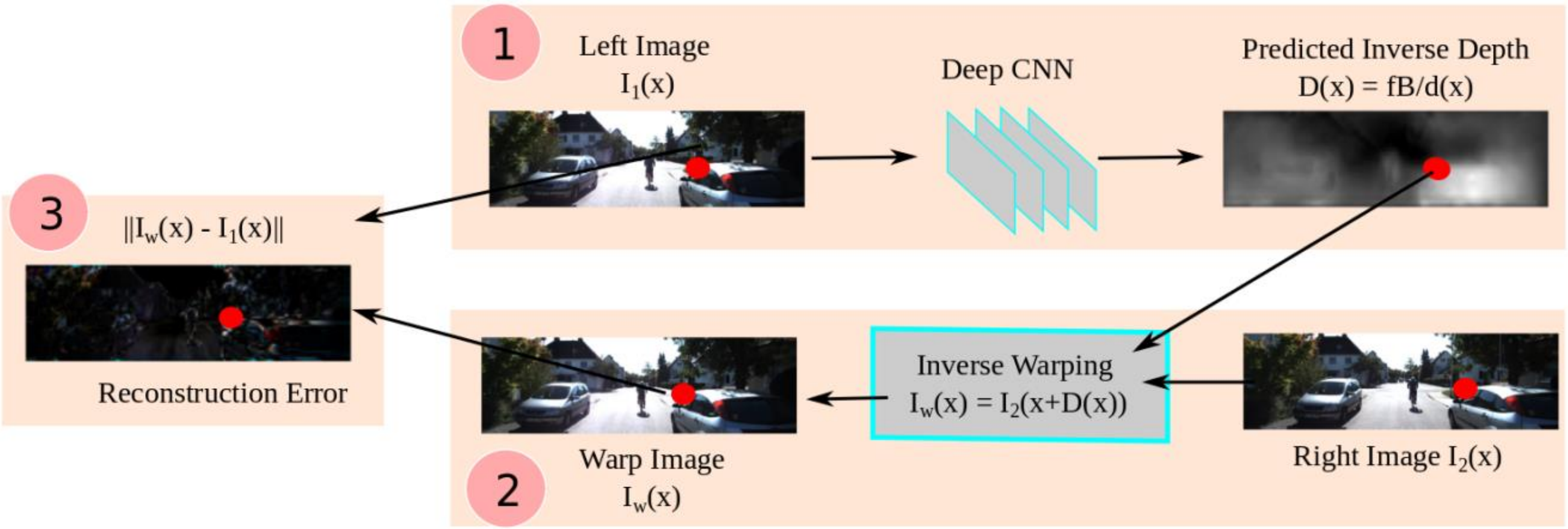


disparity

  - 1. Search for correspondences.
  - 2. Camera intrinsics and relative poses.
    - Known in Stereo Matching or solved by Structure-from-Motion.
  - 3. Depth and disparity are inverse. $D(x) = fB/d(x)$

1. 单目无监督深度估计原理

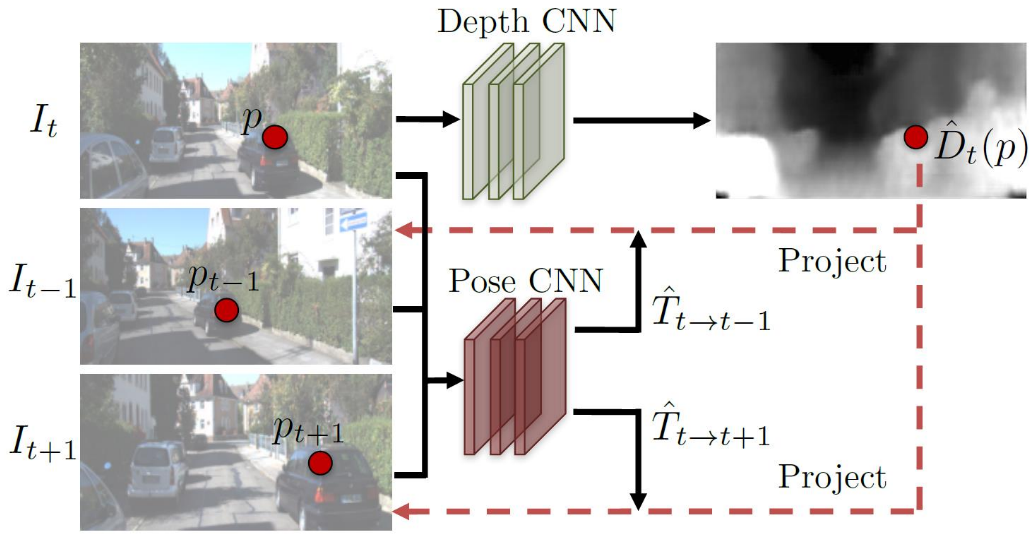
- Stereo Pair (Grag et al, ECCV 2016)



  - 1. Absolute scale.
  - 2. Occlusion issue.

1. 单目无监督深度估计原理

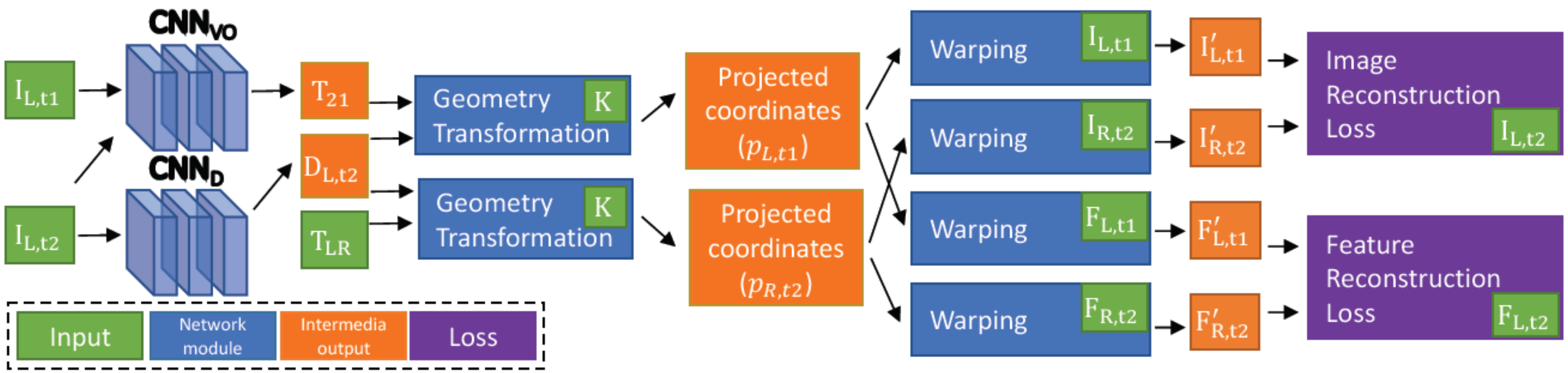
- Monocular Video (SfMLearner, CVPR 2017)



  - 1. No absolute scale. (scale ambiguity)
  - 2. Occlusion issue.
  - 3. Dynamics issue.
  - 4. Scale inconsistency issue. (cannot do visual odometry)

1. 单目无监督深度估计原理

- Stereo Video (Depth-VO-Feat, CVPR 2018)



  - Absolute scale.
  - Can do Visual Odometry.
  - Occlusion issue.
  - Dynamics issue.

1. 单目无监督深度估计原理

- Depth Results (copied from Depth-VO-Feat, CVPR 2018)

Method	Dataset	Supervision	Error metric				Accuracy metric		
			Abs Rel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Depth: cap 80m									
Train set mean	K	Depth	0.361	4.826	8.102	0.377	0.638	0.804	0.894
Eigen *et al.* [5] Fine	K	Depth	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al.* [22]	K	Depth	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Zhou *et al.* [44]	K	Mono.	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Garg *et al.* [6]	K	Stereo	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Godard *et al.* [9]	K	Stereo	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Ours (Temporal)	K	Stereo	0.144	1.391	5.869	0.241	0.803	0.928	0.969
Ours (Full-NYUv2)	K	Stereo	0.135	1.132	5.585	0.229	0.820	0.933	0.971

  - "Garg et al. [6]" stands for (Grag et al, ECCV 2016)
  - "Zhou et al. [44]" stands for (SfMLearner, CVPR 2017)
  - "Ours" stand for (Depth-VO-Feat, CVPR 2018)

1. 单目无监督深度估计原理

- Pose results (SfMLearner, CVPR 2017)

Method	**Seq.** 09	**Seq.** 10
ORB-SLAM (full)	$\mathbf{0.014 \pm 0.008}$	$\mathbf{0.012 \pm 0.011}$
ORB-SLAM (short)	0.064 ± 0.141	0.064 ± 0.130
Mean Odom.	0.032 ± 0.026	0.028 ± 0.023
Ours	$\mathbf{0.021 \pm 0.017}$	$\mathbf{0.020 \pm 0.015}$

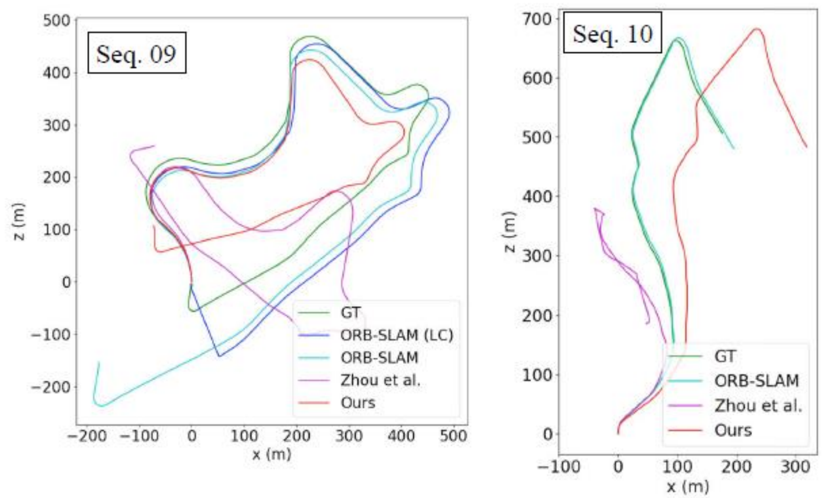
Table 3. Absolute Trajectory Error (ATE) on the KITTI odometry split averaged over all 5-frame snippets (lower is better). Our method outperforms baselines with the same input setting, but falls short of `ORB-SLAM (full)` that uses strictly more data.

1. 单目无监督深度估计原理

- Visual Odometry Results (Depth-VO-Feat, CVPR 2018)

Method	Seq. 09		Seq. 10	
	$t_{err}(\%)$	$r_{err}(°/100m)$	$t_{err}(\%)$	$r_{err}(°/100m)$
ORB-SLAM (LC) [26]	16.23	1.36	/	/
ORB-SLAM [26]	15.30	0.26	3.68	0.48
Zhou *et al.*[44]	17.84	6.78	37.91	17.78
Ours (Temporal)	11.93	3.91	12.45	3.46
Ours (Full-NYUv2)	11.92	3.60	12.62	3.43

Table 1. Visual odometry result evaluated on Sequence 09, 10 of KITTI Odometry dataset. t_{err} is average translational drift error. r_{err} is average rotational drift error.



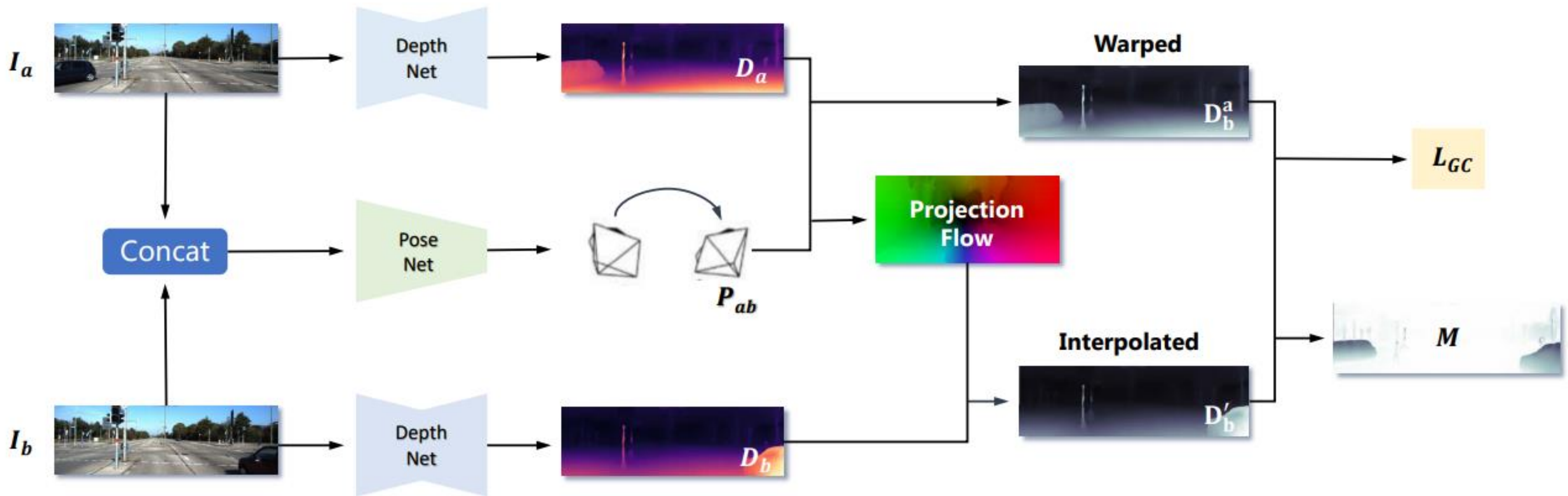
2. 输出尺度不一致问题

- 现象与影响
  - Predict depth and pose with varying scales on a sequence
  - Depth cannot be fused together for mapping
  - Poses cannot be concatenated for camera localization

- 造成问题的原因
  - Scale ambiguity
  - Photometric loss is scale-invariant
  - Training samples are independently processed.

3. 我们的解决方案

- SC-SfMLearner, NeurIPS 2019
  - Geometry Consistency Loss (for scale consistency)
  - Self Discovered Mask (for handling occlusion and dynamics)



3. 我们的解决方案

- SC-SfMLearner, NeurIPS 2019
  - Relative depth error

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

  - Geometry Consistency Loss

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

  - Self Discovered Mask

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

3. 我们的解决方案

- SC-SfMLearner, NeurIPS 2019
  - Depth and Mask Visualization

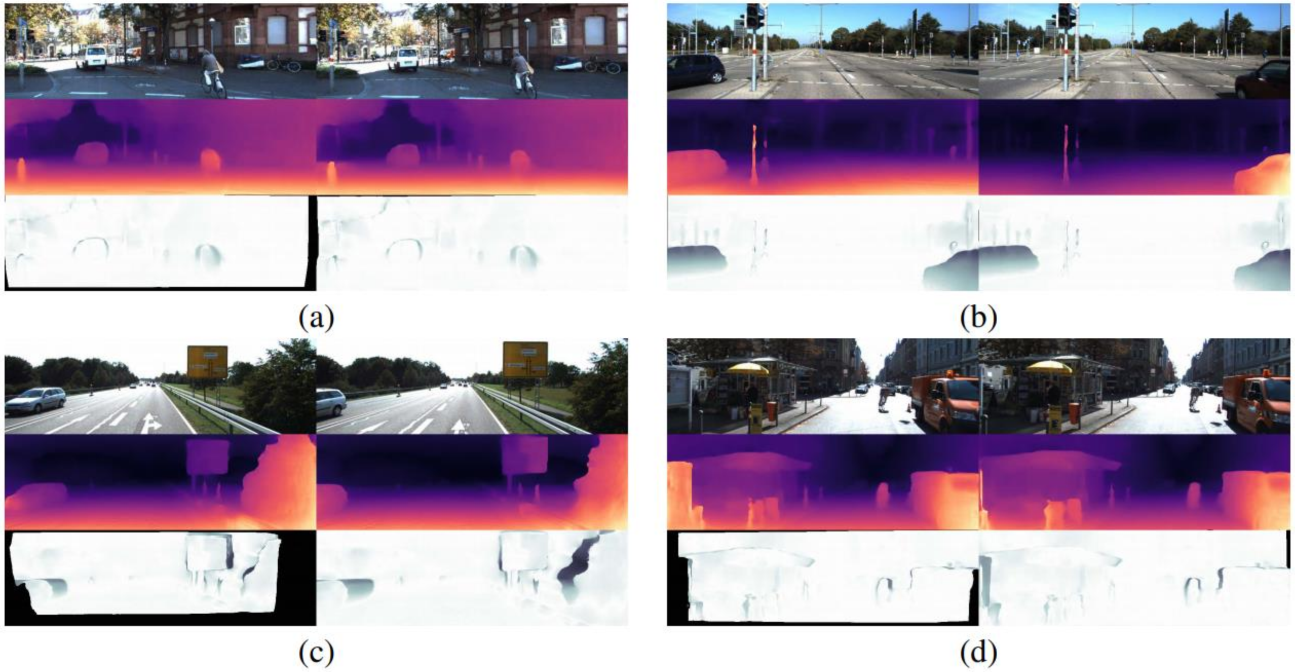


Figure 2: Visual results. Top to bottom: sample image, estimated depth, self-discovered mask. The proposed mask can effectively identify occlusions and moving objects.

3. 我们的解决方案

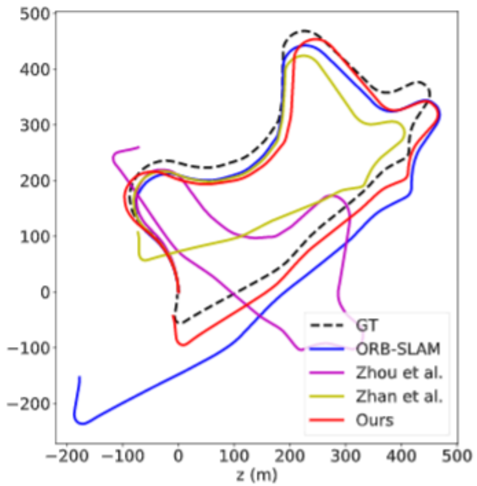
- SC-SfMLearner, NeurIPS 2019
  - Depth results

Methods	Dataset	Error ↓				Accuracy ↑		
		AbsRel	SqRel	RMS	RMSlog	< 1.25	$< 1.25^2$	$< 1.25^3$
Eigen et al. [4]	K (D)	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu et al. [5]	K (D)	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Garg et al. [21]	K (B)	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Kuznietsov et al. [18]	K (B+D)	**0.113**	**0.741**	**4.621**	**0.189**	**0.862**	**0.960**	**0.986**
Godard et al. [22]	K (B)	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard et al. [22]	CS+K (B)	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Zhan et al. [17]	K (B)	0.144	1.391	5.869	0.241	0.803	0.928	0.969
Zhou et al. [6]	K (M)	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Yang et al. [29] (J)	K (M)	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian et al. [7]	K (M)	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Wang et al. [16]	K (M)	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Geonet-VGG [8] (J)	K (M)	0.164	1.303	6.090	0.247	0.765	0.919	0.968
Geonet-Resnet [8] (J)	K (M)	0.155	1.296	5.857	0.233	0.793	0.931	0.973
DF-Net [9] (J)	K (M)	0.150	1.124	5.507	0.223	0.806	0.933	0.973
CC [10] (J)	K (M)	0.140	**1.070**	**5.326**	**0.217**	0.826	0.941	**0.975**
Ours	K (M)	**0.137**	1.089	5.439	**0.217**	**0.830**	**0.942**	**0.975**
Zhou et al. [6]	CS+K (M)	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Yang et al. [29] (J)	CS+K (M)	0.165	1.360	6.641	0.248	0.750	0.914	0.969
Mahjourian et al. [7]	CS+K (M)	0.159	1.231	5.912	0.243	0.784	0.923	0.970
Wang et al. [16]	CS+K (M)	0.148	1.187	5.496	0.226	0.812	0.938	0.975
Geonet-Resnet [8] (J)	CS+K (M)	0.153	1.328	5.737	0.232	0.802	0.934	0.972
DF-Net [9] (J)	CS+K (M)	0.146	1.182	5.215	0.213	0.818	0.943	**0.978**
CC [10] (J)	CS+K (M)	0.139	**1.032**	**5.199**	0.213	0.827	0.943	0.977
Ours	CS+K (M)	**0.128**	1.047	5.234	**0.208**	**0.846**	**0.947**	0.976

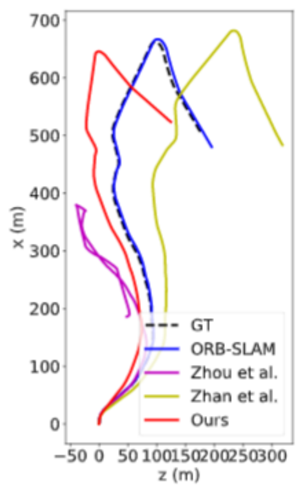
3. 我们的解决方案

- SC-SfMLearner, NeurIPS 2019
  - Visual Odometry Results

Methods	Seq. 09		Seq. 10	
	t_{err} (%)	r_{err} (°/100m)	t_{err} (%)	r_{err} (°/100m)
ORB-SLAM [11]	15.30	0.26	3.68	0.48
Zhou et al. [6]	17.84	6.78	37.91	17.78
Zhan et al. [17]	11.93	3.91	12.45	**3.46**
Ours (K)	11.25	5.85	10.10	8.56
Ours (CS+K)	**8.23**	**3.83**	**9.96**	6.90



(a) sequence 09



(b) sequence 10

3. 我们的解决方案

- SC-SfMLearner, NeurIPS 2019
  - Depth估计存在的问题
    - Although consistent, but the scale is still unknown
  - Visual Odometry存在的问题
    - Lack of multi-view optimization
    - Heavy drifts in long videos

4. 用输出尺度一致的深度做视觉SLAM

- Pseudo-RGBD SLAM (extension to SC-SfMLearner)
  - 系统介绍
    - 使用ORB-SLAM2 （RGB-D）作为框架
    - 用网络估计的depth作为输入
    - 用网络估计的pose来作为tracking的初值
  - 系统优缺点
    - 解决了单目SLAM中的初始化问题
    - 利用bundle adjustment实现multi-frame optimization
    - 利用loop closing实现drift correction
    - 创建Dense地图

4. 用输出尺度一致的深度做视觉SLAM

- Pseudo-RGBD SLAM (extension to SC-SfMLearner)
  - Visual Odometry Results on KITTI

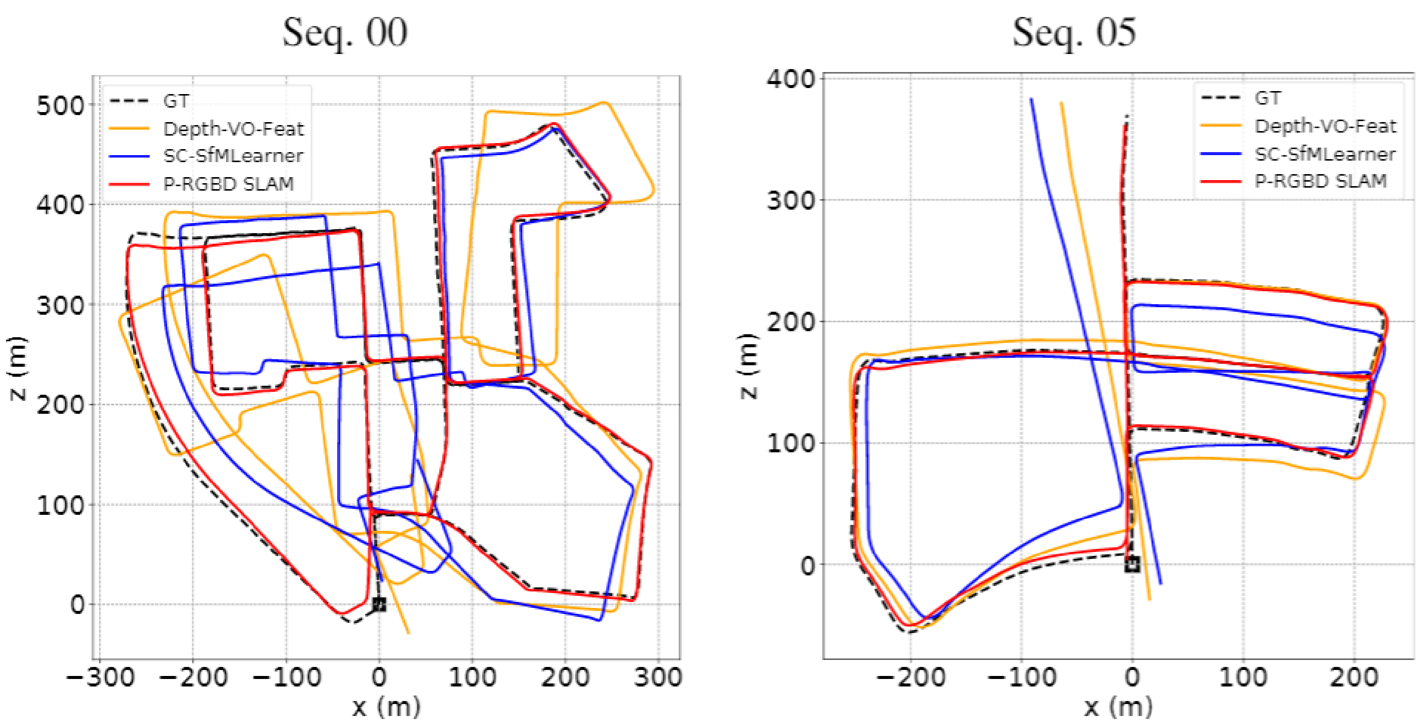
    Models are trained on (00-08). * denotes sequences that contain loops. G denotes geometric methods, M / B denotes models trained on Monocular / Binocular videos.

    7-DoF optimization is used for monocular methods (M).

Seq	VISO2-M [25] (G)		Depth-VO-Feat [9] (B)		SfMLearner [1] (M)		DW [22] (M)		SC-SfMLearner Ours (M)		P-RGBD SLAM Ours (M+G)	
	t_{err}	r_{err}	t_{err}	r_{err}	t_{err}	r_{err}	t_{err}	r_{err}	t_{err}	r_{err}	t_{err}	r_{err}
00*	10.53	2.73	6.23	2.44	21.32	6.19	11.61	3.85	6.95	2.92	**2.15**	**0.81**
02*	18.71	1.19	6.59	2.26	24.10	4.18	6.99	2.23	6.20	2.72	**2.32**	**0.73**
03	30.21	2.21	15.76	10.62	12.56	4.52	13.26	7.61	5.11	4.20	**2.47**	**1.26**
04	34.05	1.78	**3.14**	2.02	4.32	3.28	6.30	3.18	4.13	2.77	3.58	**1.84**
05*	13.16	3.65	4.94	2.34	12.99	4.66	14.36	4.22	5.91	2.96	**1.67**	**0.50**
06*	17.69	1.93	5.80	2.06	15.55	5.58	4.92	1.38	5.98	2.83	**2.48**	**1.14**
07*	10.80	4.67	6.49	3.56	12.61	6.31	14.27	9.08	7.57	3.96	**0.73**	**0.47**
08	13.85	2.52	5.45	2.39	10.66	3.75	8.09	1.83	6.84	3.10	**4.39**	**1.37**
09*	18.06	1.25	11.89	3.60	11.32	4.07	7.89	2.22	7.31	3.05	**5.08**	**1.05**
10	26.10	3.26	12.82	3.41	15.25	4.06	13.18	2.54	7.79	4.90	**4.32**	**2.34**
Avg	15.25	2.34	6.71	2.60	17.27	4.73	9.69	3.00	6.60	3.02	**2.85**	**0.95**

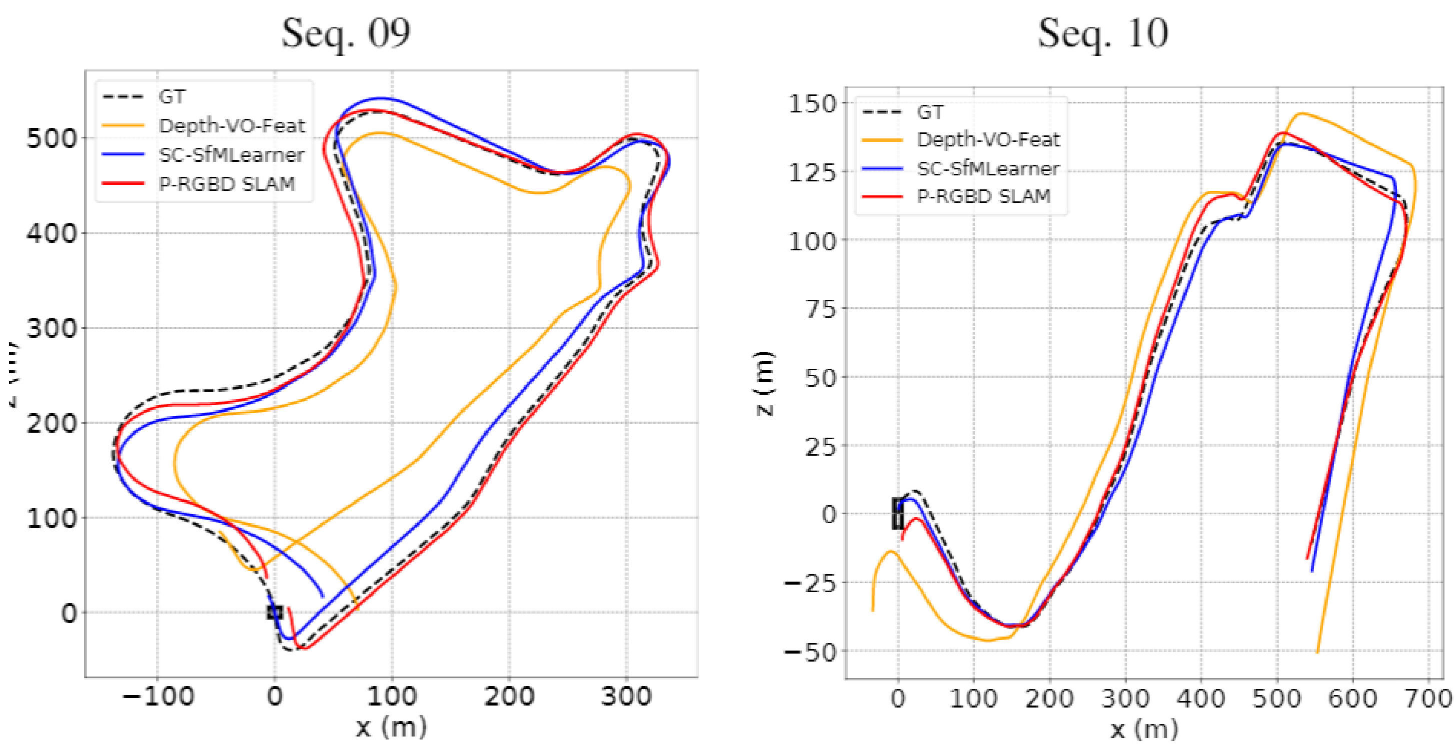
4. 用输出尺度一致的深度做视觉SLAM

- Pseudo-RGBD SLAM (extension to SC-SfMLearner)
  - Visualization of trajectory (loop detected sequences)



4. 用输出尺度一致的深度做视觉SLAM

- Pseudo-RGBD SLAM (extension to SC-SfMLearner)
  - Visualization of trajectory (no loop or loop not detected)



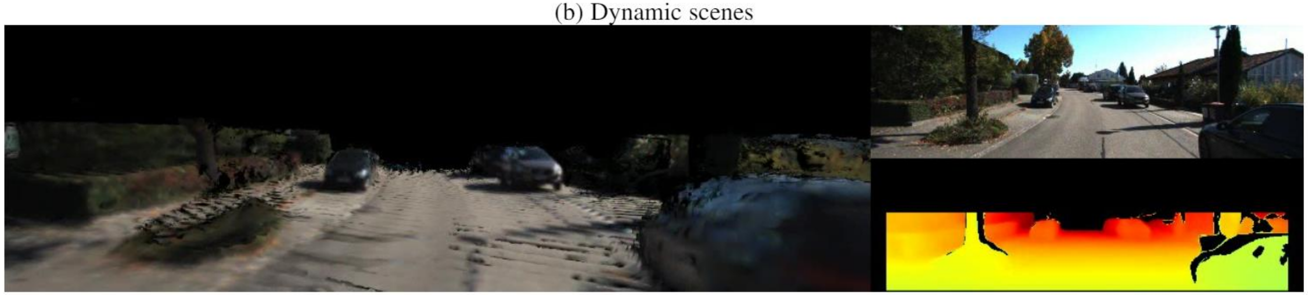
5. 三维重构Demo

- Screen shot

  Left is the reconstructed 3D model. Top right is input RGB image, Bottom right is the estimated depth map (with our mask).

(a) Static scenes

(b) Dynamic scenes



- Full Video demo link: [Here]

相关链接

- 论文地址：
  - https://arxiv.org/abs/1908.10553
- 代码地址：
  - https://github.com/JiawangBian/SC-SfMLearner-Release
- KITTI VO Evaluation Code (python):
  - https://github.com/Huangying-Zhan/kitti-odom-eval

Q & A

SUB CRUCE LUMEN

THE UNIVERSITY of ADELAIDE