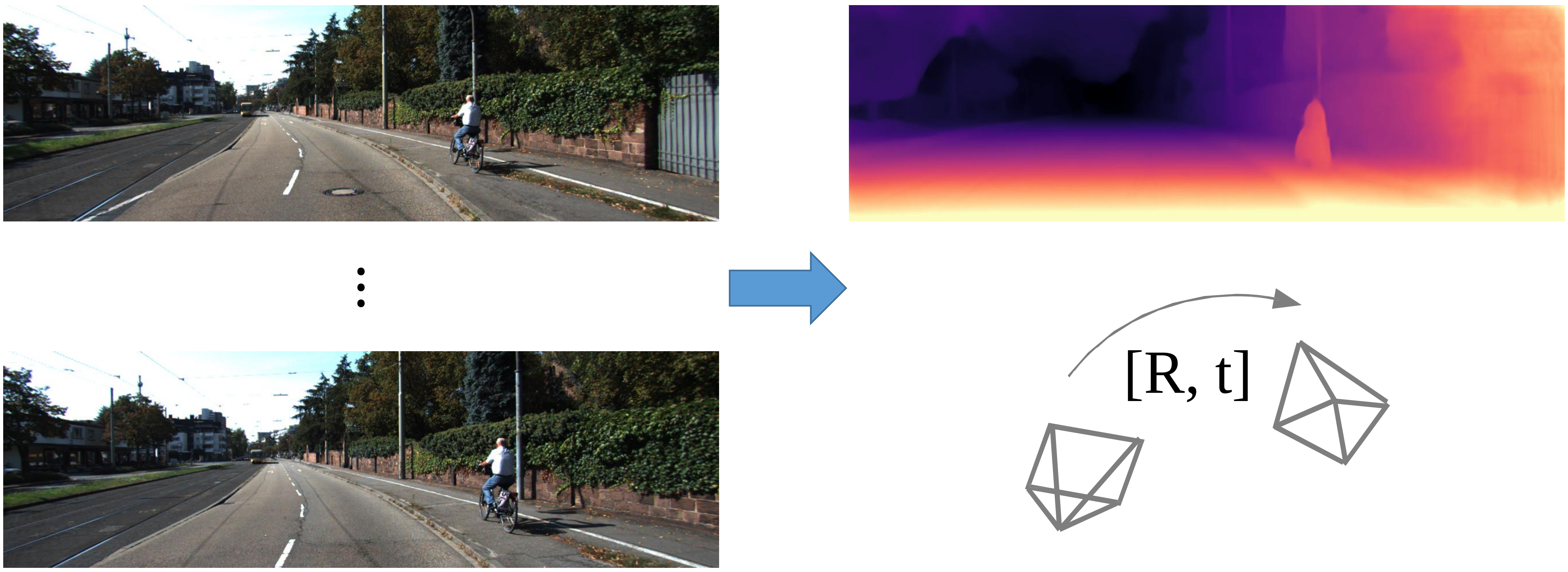
Towards Better Generalization: Joint Depth-Pose Learning without PoseNet

Wang Zhao, Shaohui Liu, Yezhi Shu, Yong-Jin Liu

Tsinghua University

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

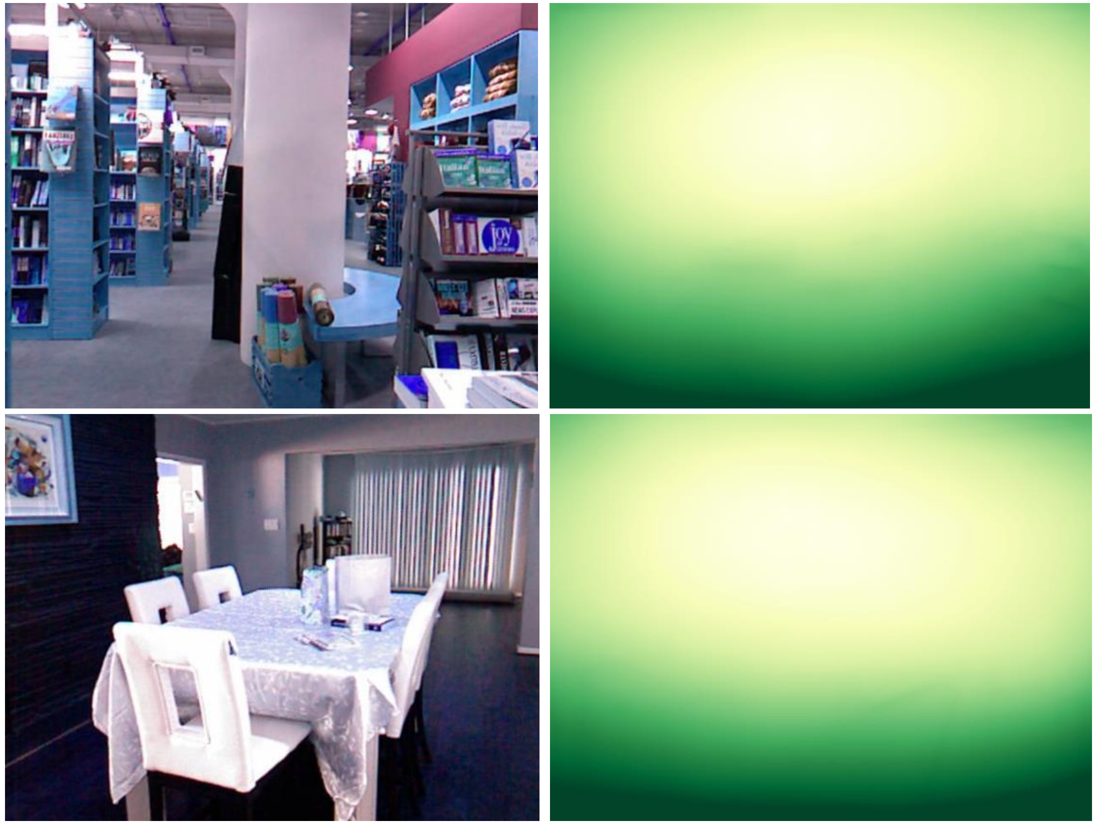
Monocular Depth-Pose Prediction



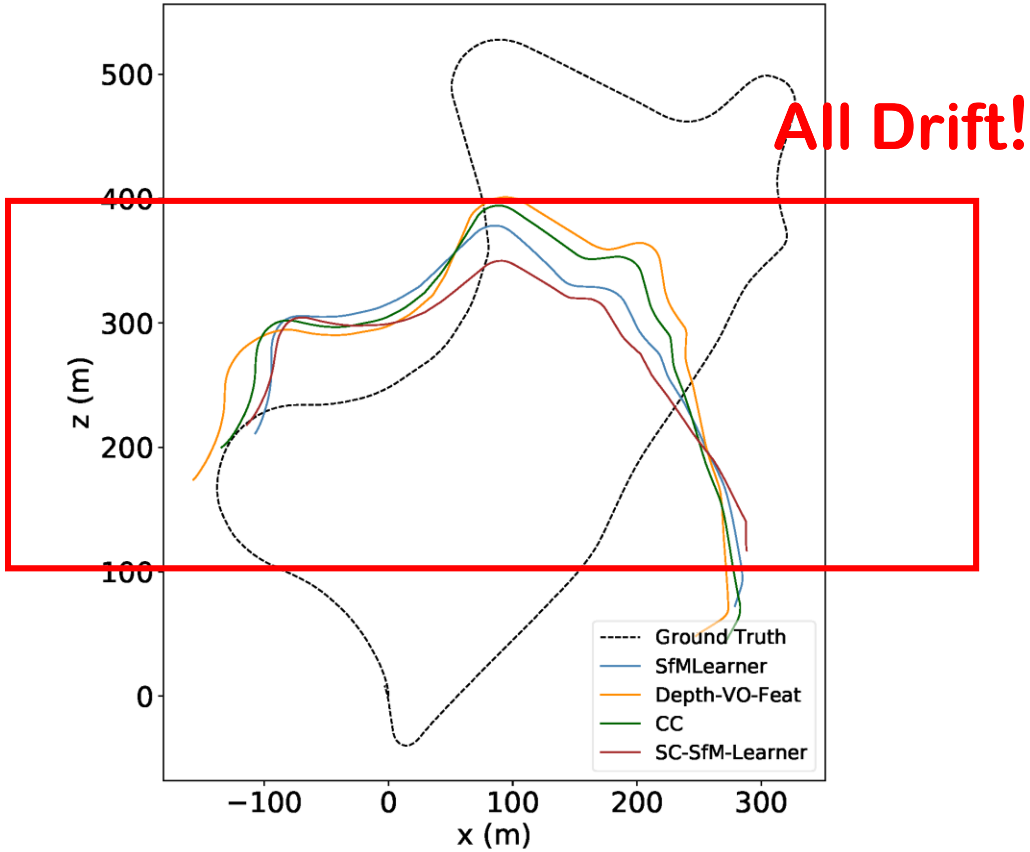
RGB

Depth and Pose

PoseNet Fails to Generalize!

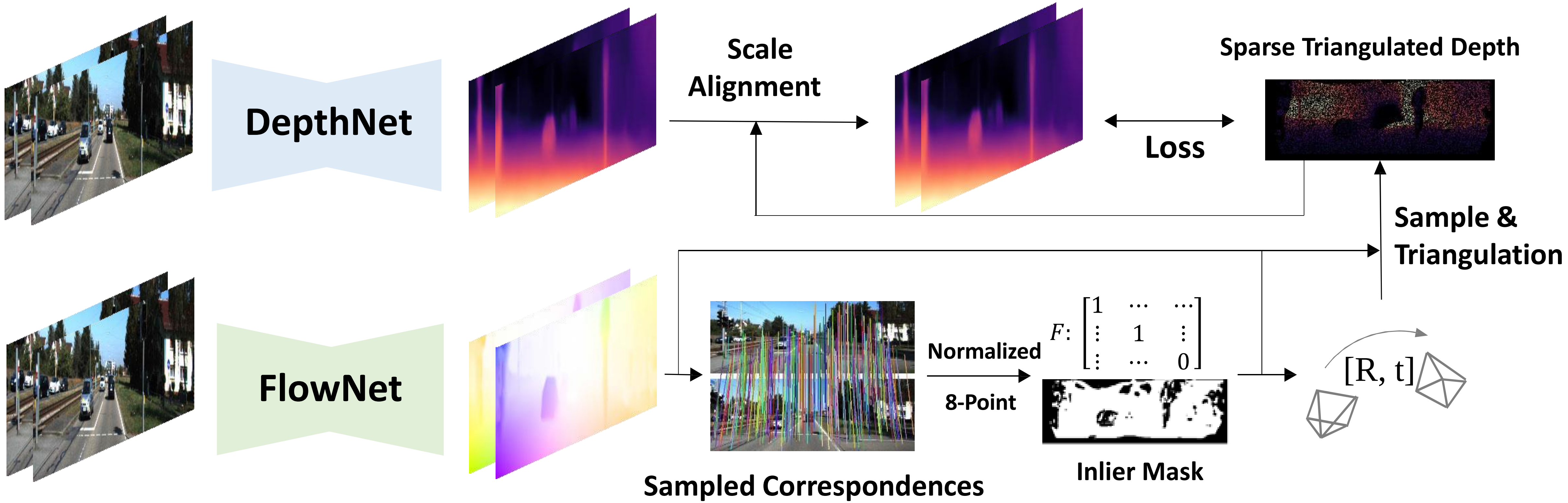


Depth estimation in Indoor environments with complex camera motions and low texture



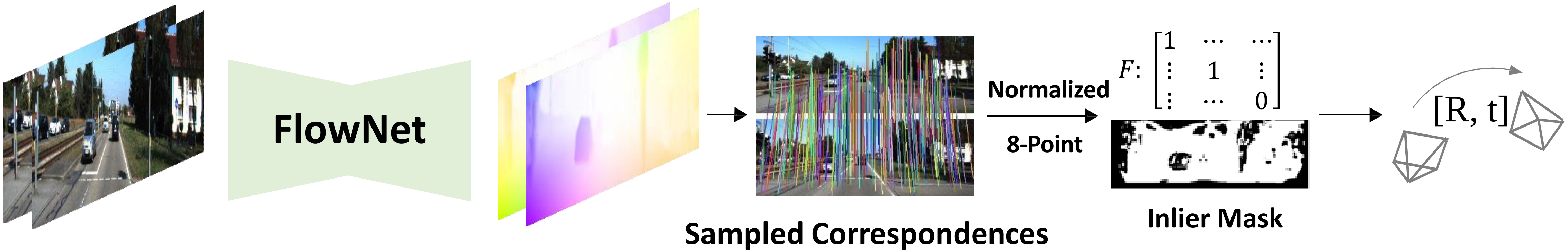
Visual Odometry with Unseen Camera Ego-motions

Joint Learning without PoseNet



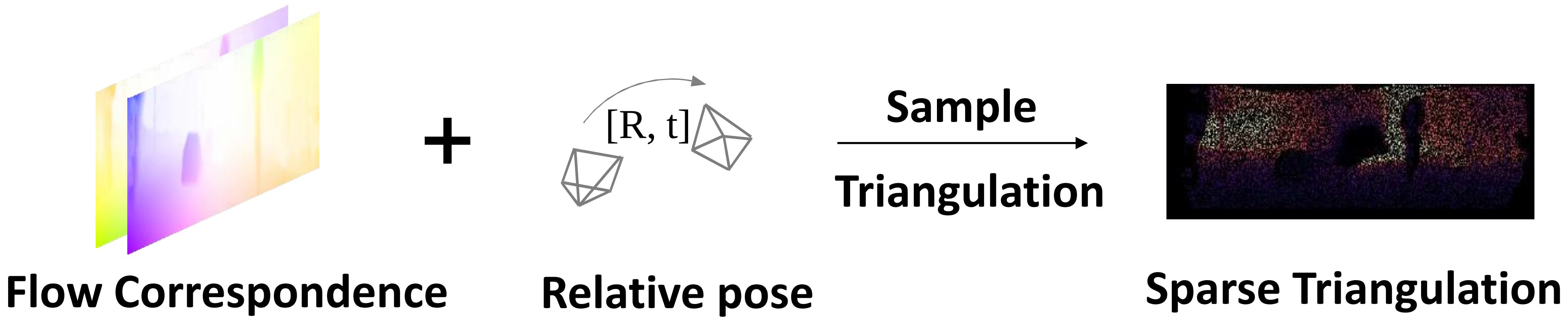
Built on top of two-frame structure-from-motion

Joint Learning without PoseNet



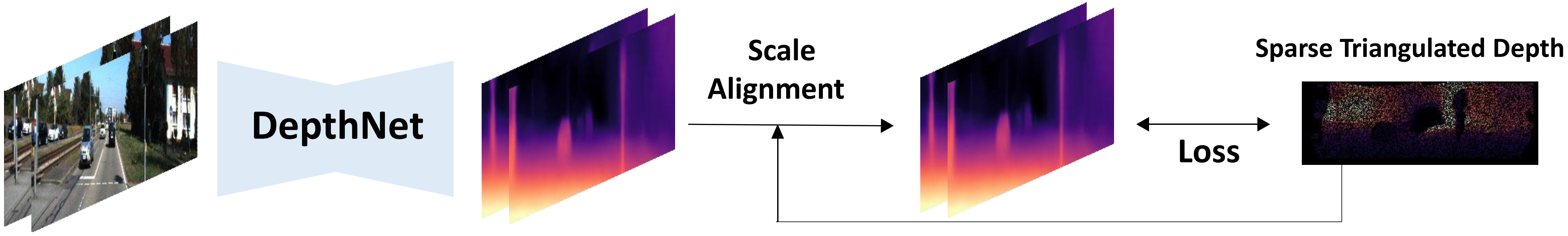
- Correspondences are sampled based on the occlusion mask and the forward-backward consistency score produced by the optical flow network .
- 8-Point algorithm is implemented in RANSAC loop to robustly recover the relative pose.
- Epipolar distance (Inlier mask) is calculated and used to further filter out the incorrect matchings and non-rigid objects.

Joint Learning without PoseNet



- We sample 6k matches from flow to triangulate, according to the occlusion mask, forward-backward score, and the inlier mask.
- We use mid-point triangulation for its convenience and it’s naturally differentiable.
- A match is abandoned if the angle between two rays is too small.

Joint Learning without PoseNet



- Predicted depth is aligned with triangulation depth map to have a consistent scale.
- Triangulation loss, depth re-projection loss and the depth smoothness loss are used to supervise the depth-net.

Scale Disentanglement

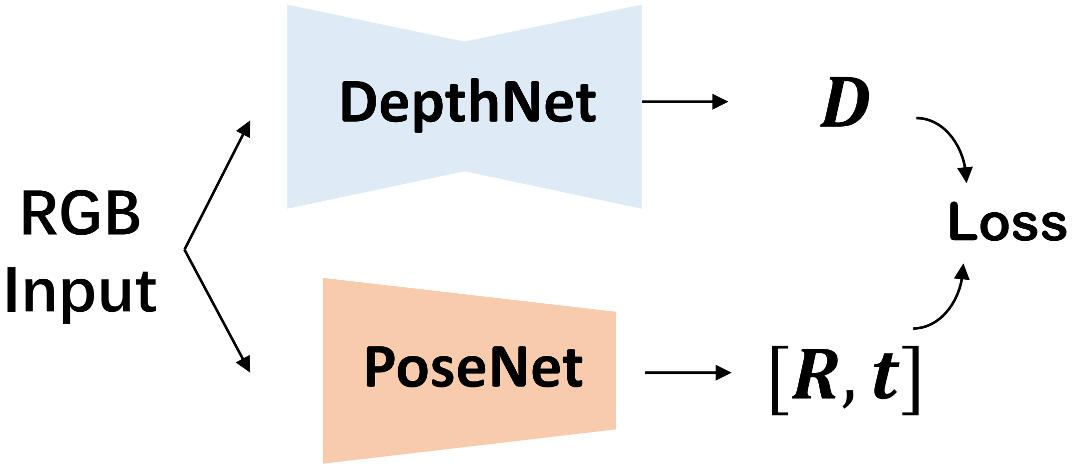
1. The translation value t of estimated pose $[R, t]$ from monocular video is up-to-scale!

2. Monocular depth prediction D from network has a learnt scale.

3. Joint training losses require a consistent scale across learnt depth and pose.

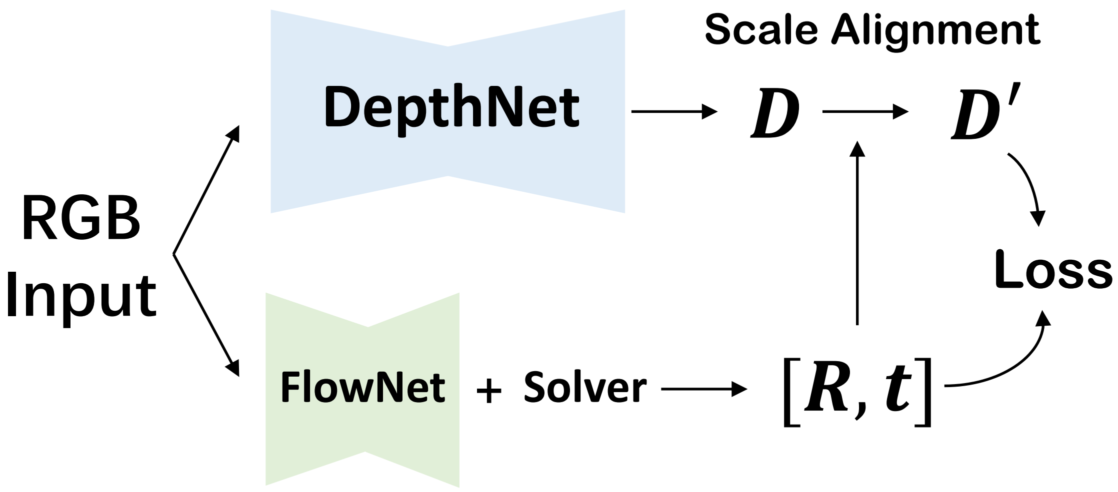
Scale Disentanglement

PoseNet-based learning system



PoseNet needs to learn a translation scale consistent with DepthNet

Our system



No need for network to learn a translation scale consistent with DepthNet

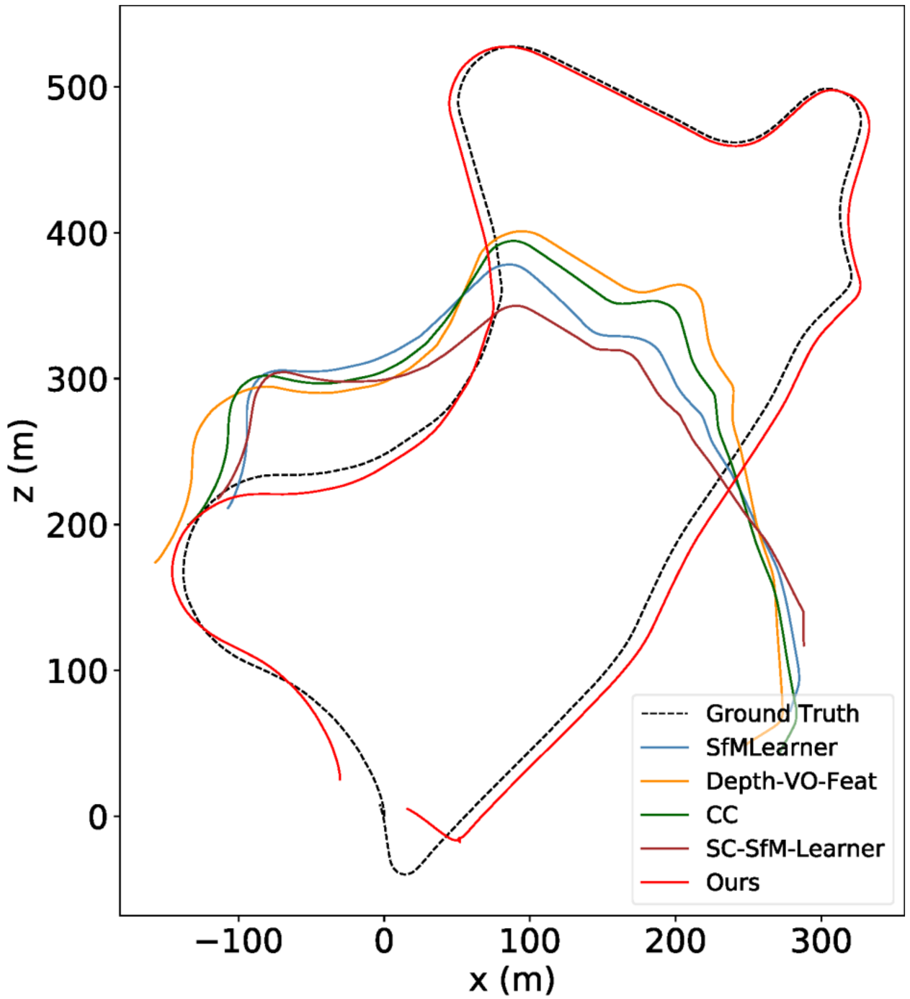
Quantitative Results on KITTI dataset

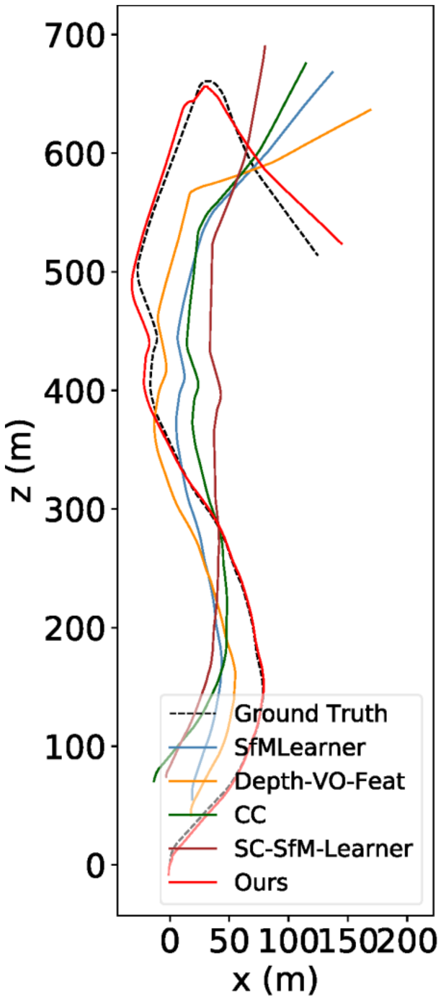
Method	Error				Accuracy, δ		
	AbsRel	SqRel	RMS	RMSlog	<1.25	$<1.25^2$	$<1.25^3$
Zhou *et al.* [66]	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Mahjourian *et al.* [35]	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Geonet [61]	0.155	1.296	5.857	0.233	0.793	0.931	0.973
DDVO [54]	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [67]	0.150	1.124	5.507	0.223	0.806	0.933	0.973
CC [41]	0.140	1.070	5.326	0.217	0.826	0.941	0.975
EPC++ [34]	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth (-ref.) [5]	0.141	1.026	5.291	0.215	0.816	0.945	0.979
GLNet (-ref.) [6]	0.135	1.070	5.230	0.210	0.841	0.948	0.980
SC-SfMLearner [2]	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Gordon *et al.* [16]	0.128	0.959	5.230	0.212	0.845	0.947	0.976
Monodepth2 [14]	0.132	1.044	5.142	0.210	0.845	0.948	0.977
Monodepth2† [14]	0.115	0.882	4.701	0.190	**0.879**	**0.961**	0.982
Ours (w/o L_p)	0.135	0.932	5.128	0.208	0.830	0.943	0.978
Ours (w/o pretraining)	0.130	0.893	5.062	0.205	0.832	0.949	0.981
Ours†	**0.113**	**0.704**	**4.581**	**0.184**	0.871	**0.961**	**0.984**

Method	Noc	All	Fl
FlowNetS [22]	8.12	14.19	-
FlowNet2 [51]	4.93	10.06	30.37%
UnFlow [37]	-	8.10	23.27%
Back2Future [23]	-	7.04	24.21%
Geonet [61]	8.05	10.81	-
DF-Net [67]	-	8.98	26.01%
EPC++ [34]	-	5.84	-
CC [41]	-	**5.66**	20.93%
GLNet [6]	4.86	8.35	-
Ours (FlowNet-only)	4.96	8.97	25.84%
Ours	**3.60**	5.72	**18.05%**

Our method achieves state-of-the-art performances on KITTI depth and optical flow estimation.

Robustness Improved – KITTI



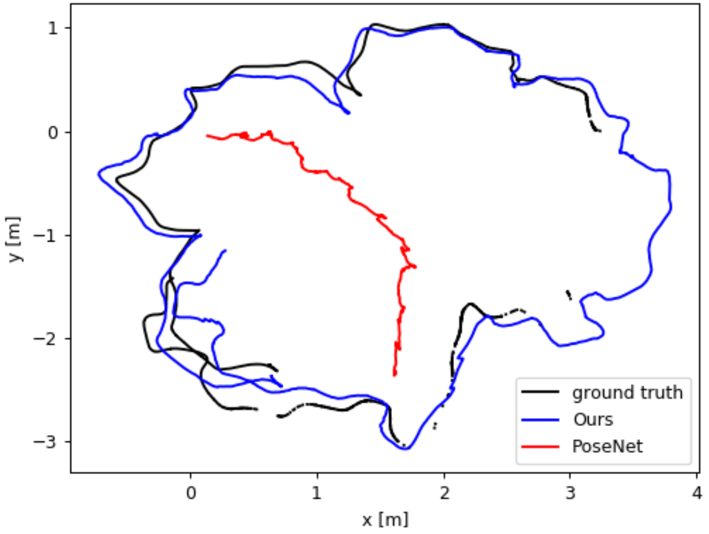


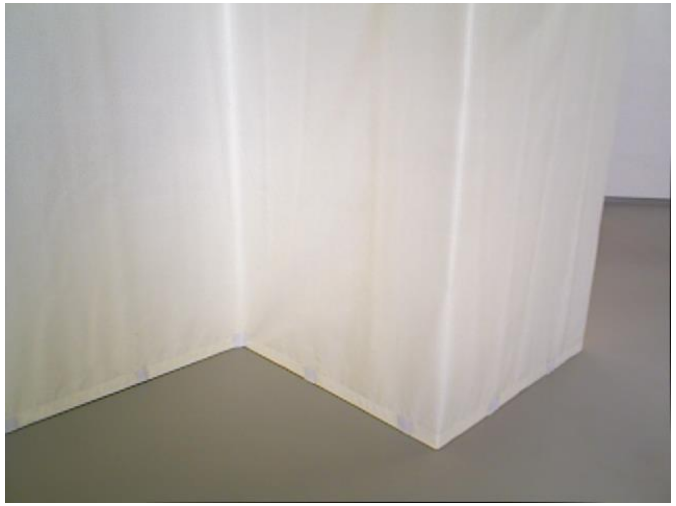
Visual Odometry with unseen camera ego-motion

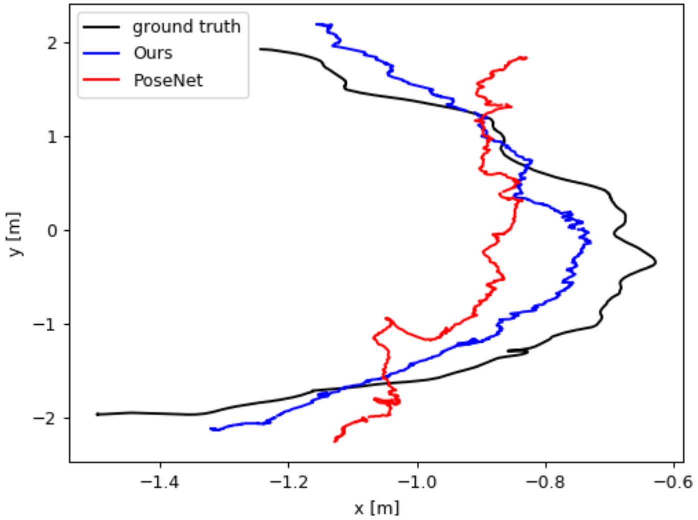
PoseNet-based ✗

Our system ✓

Robustness Improved – TUM





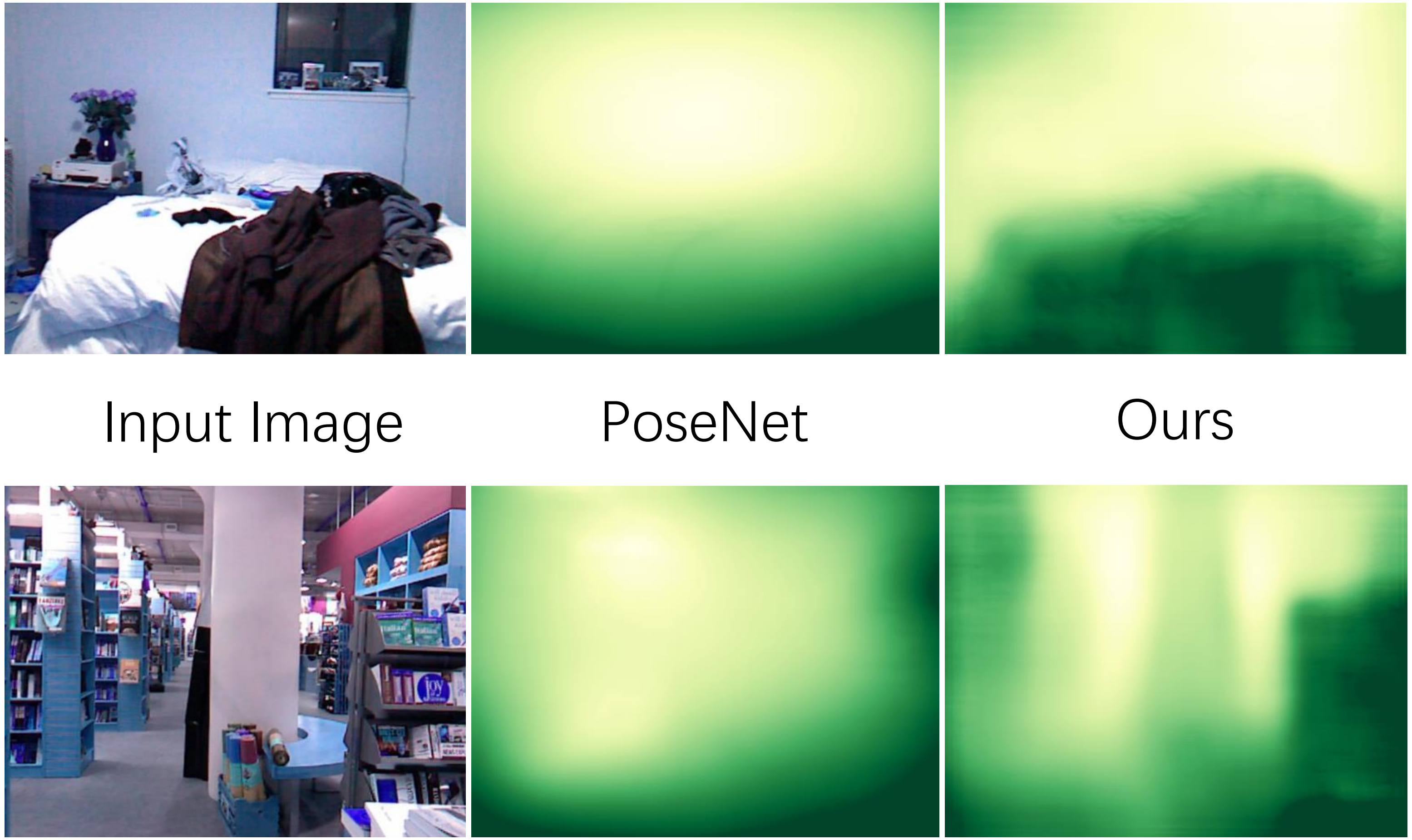


Visual Odometry with Indoor Environments

PoseNet-based ✗

Our system ✓

Robustness Improved – NYUv2



Depth Estimation in Indoor Environments

PoseNet-based

Our system

Robustness Improved – NYUv2

Method	Error			Accuracy, δ		
	rel	log10	rms	<1.25	$<1.25^2$	$<1.25^3$
Make3D [46]	0.349	-	1.214	0.447	0.745	0.897
Li *et al.* [28]	0.232	0.094	0.821	0.621	0.886	0.968
MS-CRF [58]	0.121	0.052	0.586	0.811	0.954	0.987
DORN [10]	0.115	0.051	0.509	0.828	0.965	0.992
Zhou *et al.* [64]	0.208	0.086	0.712	0.674	0.900	0.968
PoseNet	0.283	0.122	0.867	0.567	0.818	0.912
PoseNet-Flow	0.221	0.091	0.764	0.659	0.883	0.959
Ours	0.201	0.085	0.708	0.687	0.903	0.968
Ours (448×576)	**0.189**	**0.079**	**0.686**	**0.701**	**0.912**	**0.978**

Depth Estimation in Indoor Environments

PoseNet-based

Our system

Best performance on NYUv2 among unsupervised methods!

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Code and model are available here

Link: https://github.com/B1ueber2y/TrianFlow

Check our paper for more details!