Unsupervised Learning of Monocular Depth and Ego-Motion Using Multiple Masks

Shanghai Jiao Tong University, Shanghai China

Guangming Wang, Hesheng Wang, Yiling Liu and Weidong Chen

2019 IEEE International Conference on Robotics and Automation

May 20-24, 2019 Montreal, Canada

Task

- Unsupervised learning of depth and ego-motion from monocular video

Preliminary

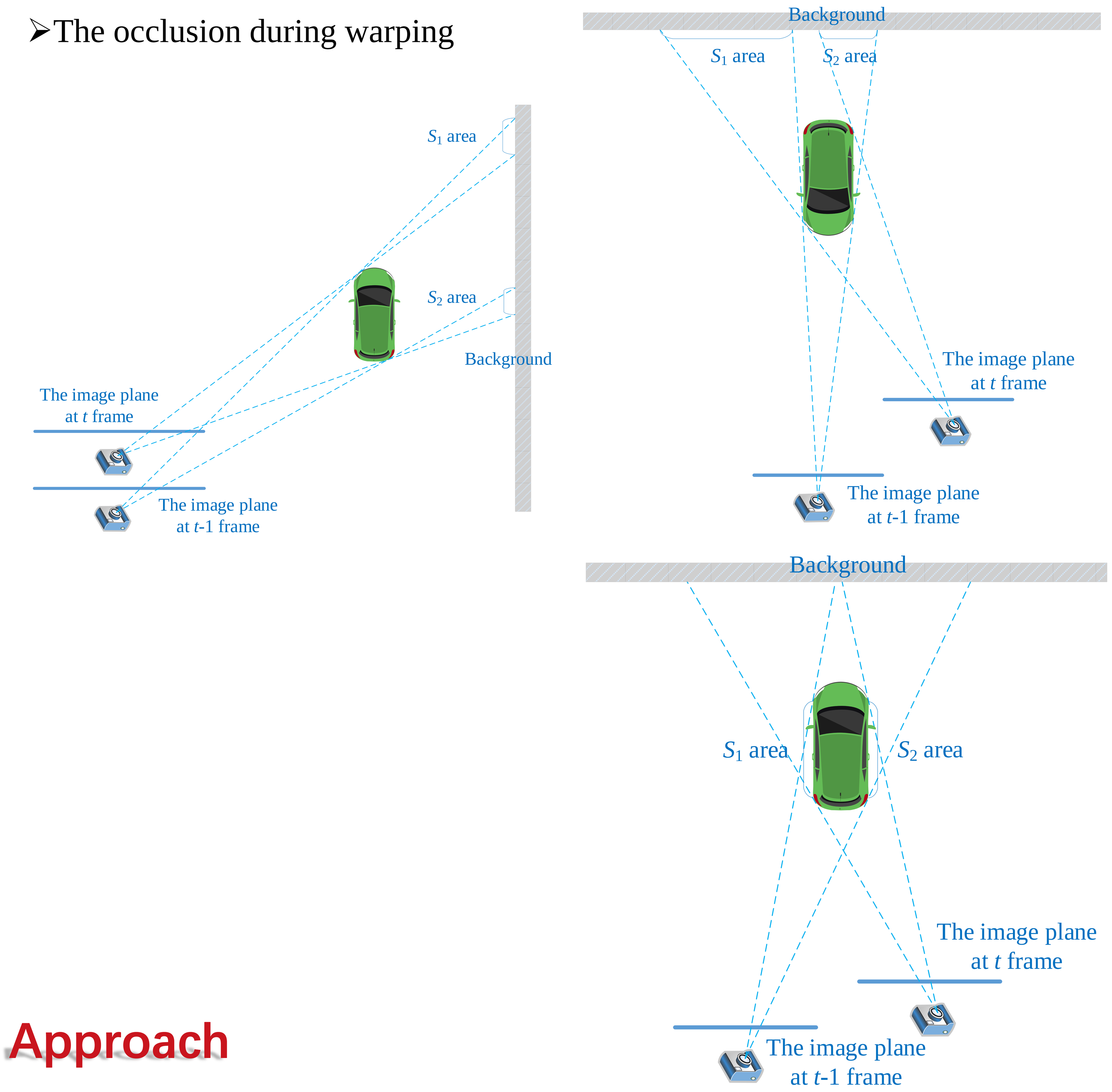
- Zhou et al. proposed to use view synthesis as supervision

Challenge

- 3D Geometric cues are modeled without any ground truth

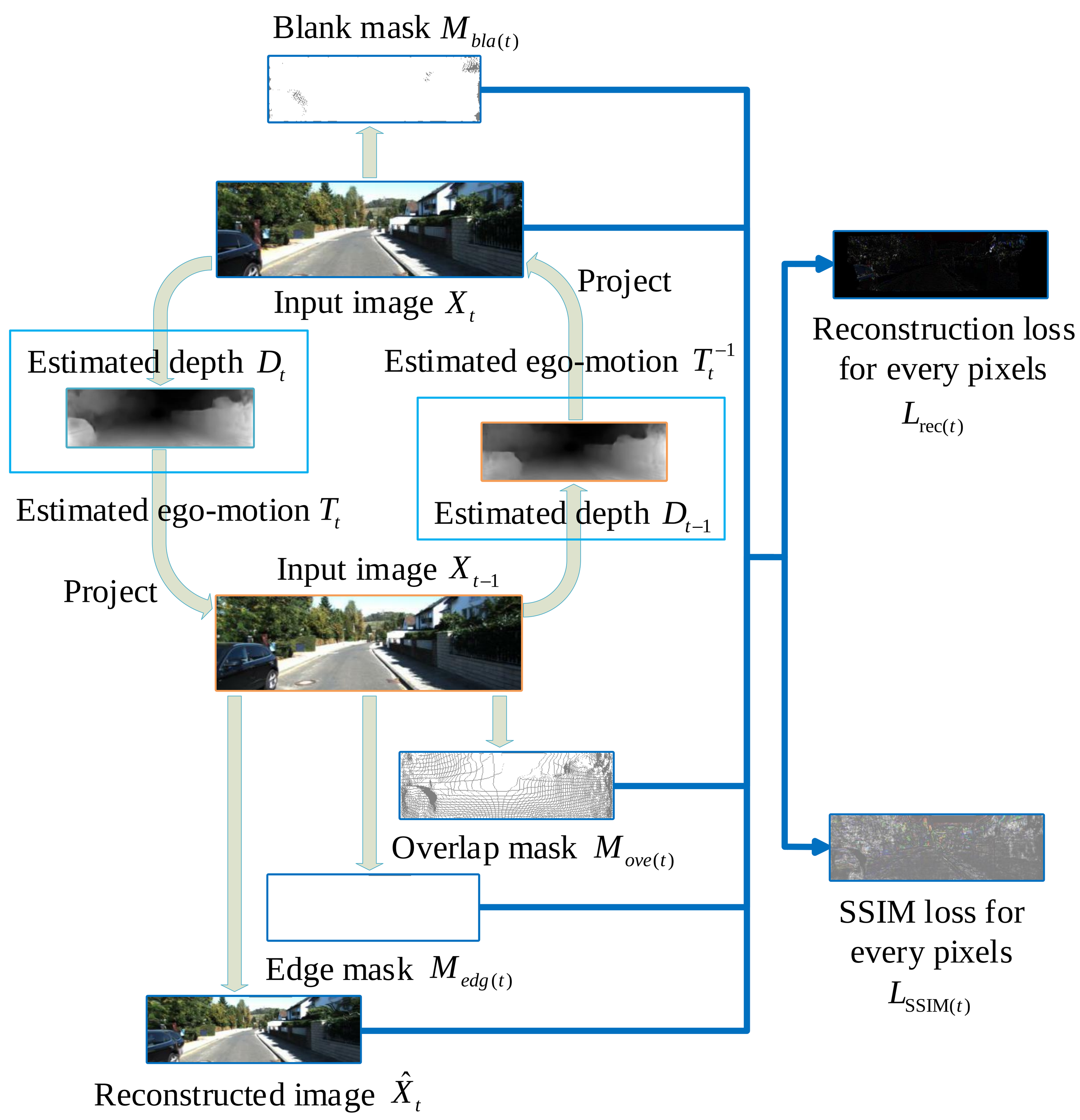
Motivation

- The occlusion during warping



Approach

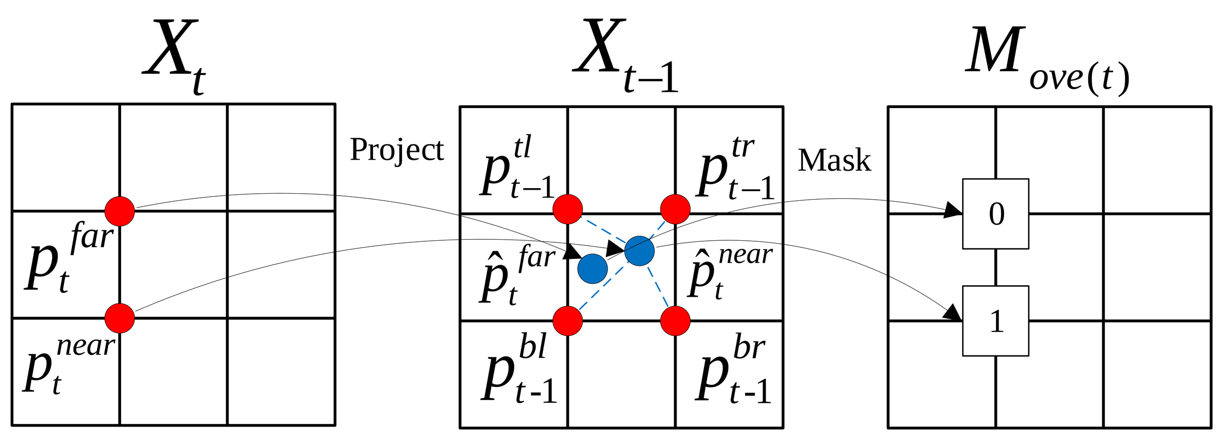
- The overview



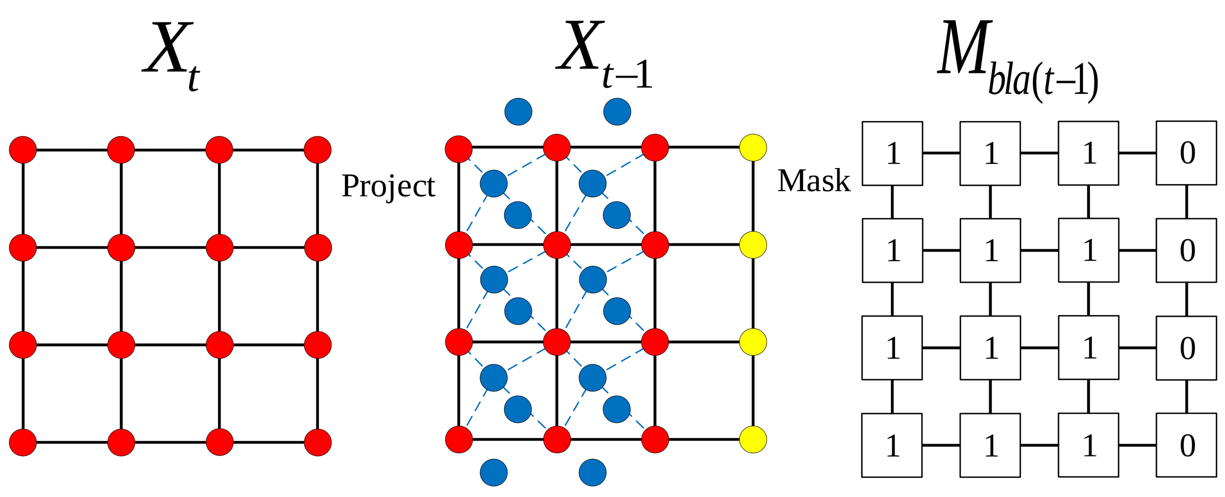
*The Figure just shows the process in one direction (t) and the other direction (t-1) is also used and similar.

- Our system uses consecutive two frames from monocular video to reconstruct the images based on the depth and ego-motion estimated by the depth estimation network and the pose estimation network, obtaining several masks at the same time.
- Next, the reconstructed image, the original image, and several masks are used to calculate the two main loss functions for backpropagation.

- The mechanism of overlap mask



- The mechanism of blank mask



- Image reconstruction loss

$$L_{\text{rec}} = \sum_{ij} \left\| (X_t^{ij} - \hat{X}_t^{ij}) M_{edg(t)}^{ij} M_{ove(t)}^{ij} M_{bla(t)}^{ij} \right\|$$

- Structural similarity loss

$$L_{\text{SSIM}} = \sum_{ij} [1 - \text{SSIM}(X_t^{ij}, \hat{X}_t^{ij})] M_{edg(t)}^{ij} M_{ove(t)}^{ij} M_{bla(t)}^{ij}$$

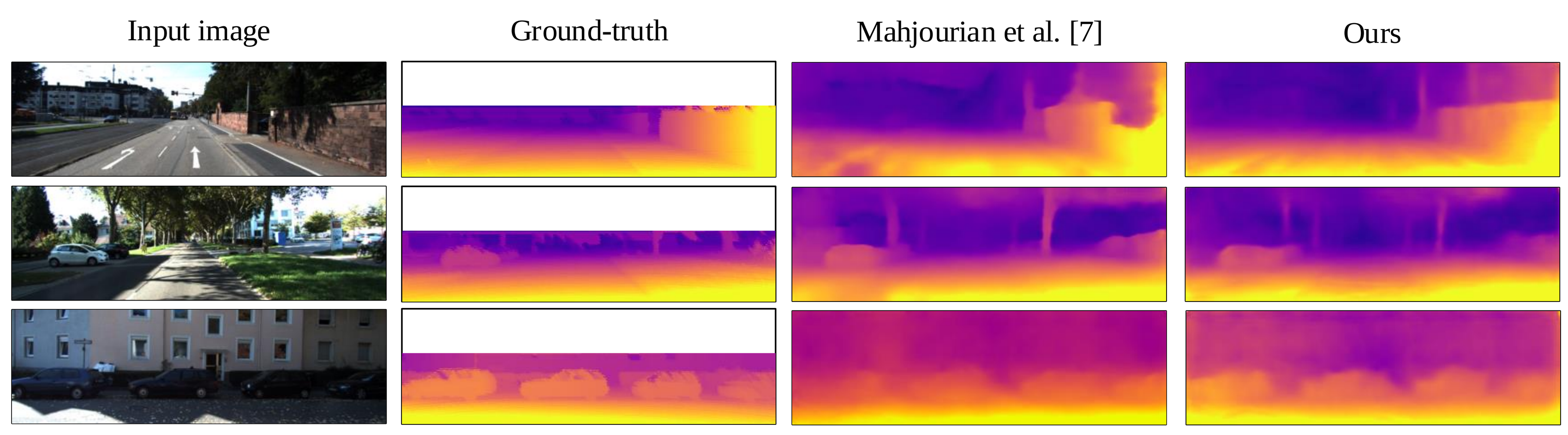
- The common smoothness loss as in [7] and [16] L_{smooth}

Results

- Depth

Method	Supervised	Dataset	Error metric: Abs Rel	Sq Rel	RMSE	RMSE log	Accuracy metric: $\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard et al. [9]	Stereo	K	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Zhan et al. [10]	Stereo	K	0.144	1.391	5.869	0.241	0.803	0.928	0.969
Zhou et al. [12]	No	K	0.208	1.768	6.856	0.283	0.678	0.885	0.957
GeoNet [16]	No	K	0.164	1.303	6.090	0.247	0.765	0.919	0.968
Mahjourian et al. [7]	No	K	0.163	1.240	6.220	0.250	0.762	0.916	0.968
LEGO [17]	No	K	0.162	1.352	6.276	0.252	-	-	-
Ours	No	K	0.158	1.277	5.858	0.233	0.785	0.929	0.973
Ours +DN*	No	K	0.154	1.163	5.700	0.229	0.792	0.932	0.974

*DN means depth normalization from [18]. The best and the second best performance in each block are highlighted in blue and green.



- Camera Pose

Method	Seq.09	Seq.10
ORB-SLAM (full)	0.014 ± 0.008	0.012 ± 0.011
ORB-SLAM (short)	0.064 ± 0.141	0.064 ± 0.130
Zhou et al. [12]	0.021 ± 0.017	0.020 ± 0.015
Mahjourian et al. [7]	0.013 ± 0.010	0.012 ± 0.011
GeoNet [16]	0.012 ± 0.007	0.012 ± 0.009
Ours	$\mathbf{0.009 \pm 0.005}$	$\mathbf{0.008 \pm 0.007}$

- Visualization of the Overlap Mask and Blank Mask