MonoIndoor: Towards Good Practice of Self-Supervised Monocular Depth Estimation for Indoor Environments

Pan Ji[*1] , Runze Li[*1,2] , Bir Bhanu[2], Yi Xu[1]
[1]OPPO US Research Center, InnoPeak Technology, Inc. [2]University of California Riverside

Appendix

Network Details

In the depth factorization module, we use the same depth network of an auto-encoder structure as in [2] to predict the relative depth, and employ a scale network consisting of an encoder and a regressor. The encoder of the scale network is shared with the depth encoder and the architecture of the scale regressor is described in Table 1. In the residual pose module, we use one pose network and one residual pose network, both of which share the same structure. The residual pose network shares parameters in the encoder with the pose network but learns independent parameters in its pose prediction head.

Table 1. Scale regressor architecture. Here **chns** is the number of ouput channels, **k** is the kernal size, **s** is the stride, **res** is the downscaling factor for each layer with respect to the input image, and **input** is the input to each layer.

Scale Regressor				
Block	**layer**	**chns-k-s**	**res**	**input**
Attention	(query) (key) (value)	(512, 1, 1) (512, 1, 1) (512, 1, 1)	32	(econv5) (econv5) (econv5)
ConvBlock1	(convs1_0) (convs1_1)	(512, 3, 1) (512, 3, 1)	32	Attention
ConvBlock2	convs2_0	1024, 1, 2	64	ConvBlock1
ConvBlock3	(Convs3_0) (Convs3_1)	(1024, 3, 1) (1024, 3, 1)	64	ConvBlock2
FC1-1024-Dropout				
FC2-1024-Dropout				
Scale Regression				

RGB-D 7-Scenes Dataset

In this section, we evaluate our **MonoIndoor** on the RGB-D 7-Scenes dataset [4] which contains several video sequences with 500-1000 frame in each sequence. All scenes are recorded using a handheld Kinect RGB-D camera at 640×480 resolution. We use the official train/test split.

*Joint first authorship. P. Ji is the corresponding author (peterji530@gmail.com). R. Li's contribution was made during an internship with OPPO US Research Center.

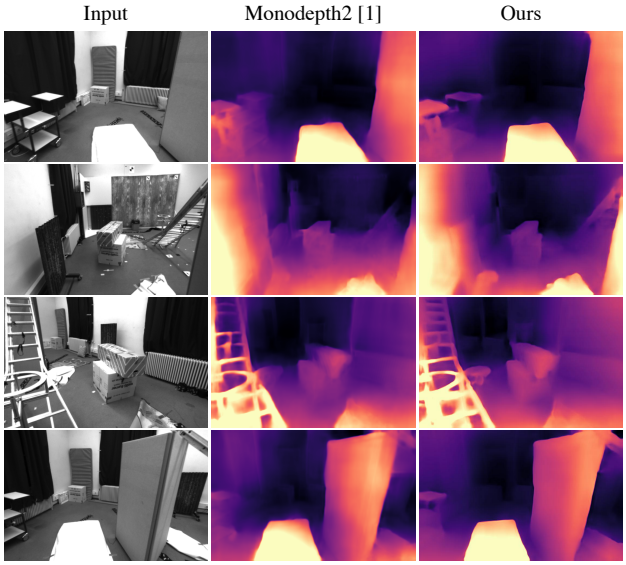

Figure 1. Additional qualitative comparison on EuRoC MAV[3].

Following [1], for training, we first pre-train our **MonoIndoor** on NYUv2 dataset, and then fine-tune the model on this dataset; for testing, we extract one image from every 30 frames. Images are resized to 320×256 during training.

We present the quantitative results of our model **MonoIndoor** and latest state-of-the-art (SOTA) self-supervised methods on 7-Scenes in Table 5. It shows that our model outperforms [1] on most scenes before and after fine-tuning, demonstrating better generalizability and capability of our model. Specifically, compared to a recent self-supervised method by Bian et al. [1], on the scene "Fire", our method reduces AbsRel by 1.2% and increases δ_1 by 2.3%, reaching an AbsRel of 7.7% and δ_1 of 93.9%; on the scene "Heads",our method reduces AbsRel by 1.8% and increases δ_1 by 2.7%, reaching an AbsRel of 10.6% and δ_1 of 88.9%.

Odometry Evaluation

In Table 3, we evaluate the proposed residual pose estimation module on the test sequences V1_03 and V2_01

Table 2. Comparison of our method to latest self-supervised methods on RGB-D 7-Scenes [4]. Best results are in **bold**.

Scenes	Bian *et al*. [1]				**MonoIndoor** (Ours)			
	Before Fine-tuning		After Fine-tuning		Before Fine-tuning		After Fine-tuning	
	AbsRel	Acc δ_1	AbsRel	Acc δ_1	AbsRel	Acc δ_1	AbsRel	Acc δ_1
Chess	0.169	0.719	0.103	0.880	0.157	0.750	**0.097**	**0.888**
Fire	0.158	0.758	0.089	0.916	0.150	0.768	**0.077**	**0.939**
Heads	0.162	0.749	0.124	0.862	0.171	0.727	**0.106**	**0.889**
Office	0.132	0.833	0.096	0.912	0.130	0.837	**0.083**	**0.934**
Pumpkin	0.117	0.857	0.083	**0.946**	0.102	0.895	**0.078**	0.945
RedKitchen	0.151	0.78	0.101	0.896	0.144	0.795	**0.094**	**0.915**
Stairs	0.162	0.765	0.106	0.855	0.155	0.753	**0.104**	**0.857**

of the EuRoC MAV [3]. We follow [6] to evaluate relative camera poses estimated by our residual pose estimation module. We use the following evaluation metrics: absolute trajectory error (ATE) which measures the root-mean square error between predicted camera poses and ground-truth, and relative pose error (RPE) which measures frame-to-frame relative pose error in meters and degrees, respectively. As shown in Table 3, on both two test sequences, compared with the baseline model Monodepth2 [2] which employs one-stage pose network, using our residual pose estimation module leads to improved relative pose estimation across all evaluation metrics. Specifically, on the sequence V1_03, the ATE by our **MonoIndoor** is significantly decreased from 0.0681 meters to 0.052 meters and PRE(°) is reduced by around half, from 1.3237°to 0.7179°.

Table 3. Odometry results on the EuRoC MAV [3] test set. Results show the average absolute trajectory error(ATE), and the relative pose error(RPE) in meters and degrees, respectively. Seq.: sequence name.

Seq.	Methods	ATE(m)	RPE(m)	RPE(°)
V1_03	Monodepth2 [2]	0.0681	0.0686	1.3237
	MonoIndoor(Ours)	**0.052**	**0.0637**	**0.7179**
V2_01	Monodepth2 [2]	0.0266	0.0199	1.1985
	MonoIndoor(Ours)	**0.0222**	**0.0109**	**1.1974**

Additional Qualitative Results

We include additional qualitative results on both the EuRoC and NYUv2 test sets in Figure 1 and Figure 2, respectively. From both figures, we can see that our models generate depth maps of higher quality.

In Figure 3, we visualize predictions qualitatively on NYUv2 by our proposed modules. We can see that each module improves the quality of depth maps and our full models produce depth maps of higher quality.

We further visualize intermediate and final synthesized views compared with the current view on NYUv2 in the Figure 4. Highlighted regions show that final synthesized views are better than the intermediate synthesized views and closer to the current view.

References

[1] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jun Chin, Chunhua Shen, and Ian Reid. Unsupervised depth learning in challenging indoor video: Weak rectification to rescue. *arXiv preprint arXiv:2006.02708*, 2020. 1, 2

[2] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019. 1, 2

[3] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 1, 2

[4] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 1, 2

[5] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 3

[6] Huangying Zhan, Chamara Saroj Weerasekera, Jia-Wang Bian, and Ian Reid. Visual odometry revisited: What should be learnt? In *ICRA*, pages 4203–4210, 2020. 2

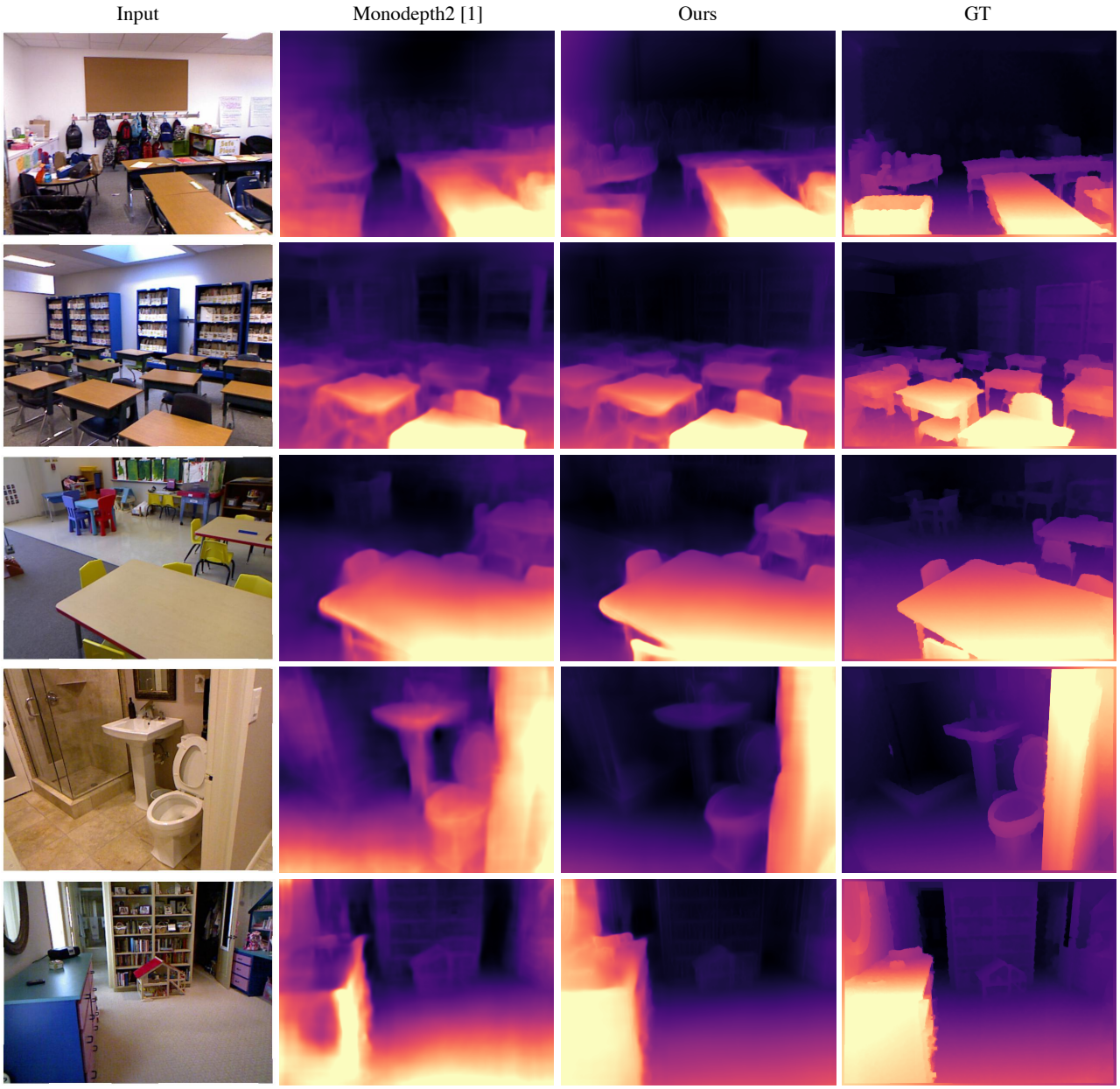

Figure 2. Additional qualitative comparison on NYUv2 [5].

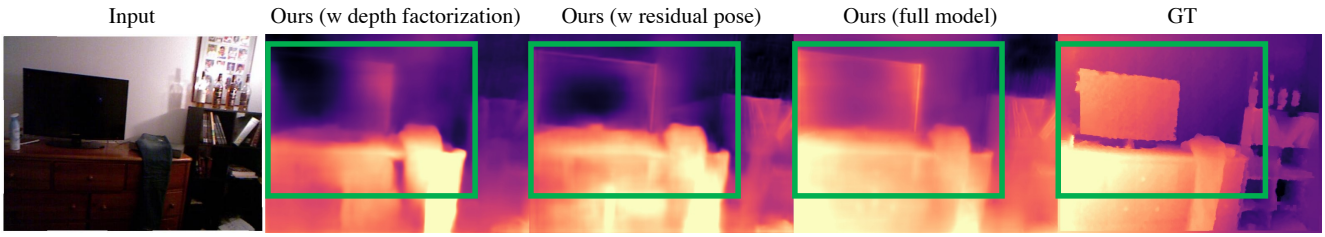

Figure 3. Qualitative ablation comparisons of depth prediction on NYUv2. Our full model with both depth factorization and residual pose modules produce better depth maps.

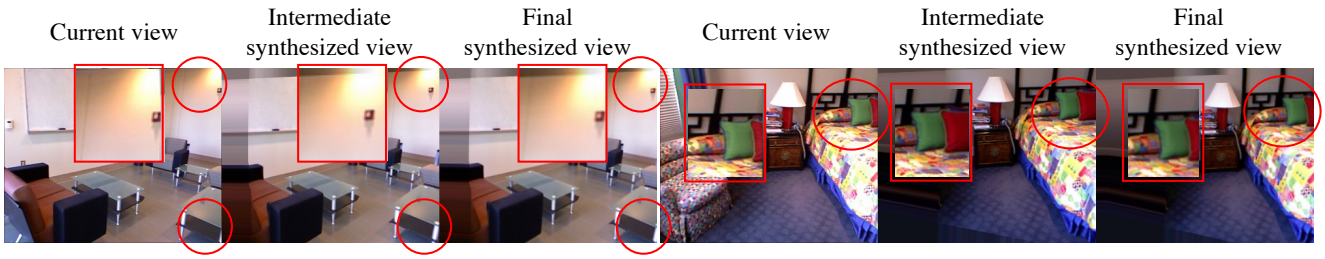

Figure 4. Intermediate synthesized views on NYUv2.