StructDepth: Leveraging the structural regularities for self-supervised indoor depth estimation (Supplementary Material)

Boying Li*, Yuan Huang*, Zeyu Liu, Danping Zou†, and Wenxian Yu
Shanghai Key Laboratory of Navigation and Location-Based Services
Shanghai Key Laboratory of Intelligent Sensing and Recognition
Shanghai Jiao Tong University

Here, we present extra experimental results, including the additional visual results of our method, the outdoor tests, and the plane quality tests.

1. Extra qualitative results

We include additional qualitative results on NYUv2, ScanNet, and InteriorNet datasets. Fig. 2 shows the 3D structure recovered from the estimated depth. Fig. 3 and Fig. 4 illustrate the results of the depth and surface normal estimation. Those results show that our method achieves more accurate depth estimation and produces more accurate 3D structures, compared with the existing methods.

2. Outdoor tests

We present the results of our method on the KITTI dataset, which is captured in outdoor scenes. We use the split composed of 44234 images as the training dataset, the same as Monodepth2[1]. We firstly detected the vanishing points on the training images and skipped 335 images which fail to detect valid vanishing points. Consequently, 39500 image sequences were used for training and 4397 image sequences for validation. Other dataset preprocessing settings are consistent with [1]. The total epoch number of training is 17 with a batch size of 16. The initial learning rate is 10^{-5} and drops to 10^{-6} after 15 epochs. Results are shown in Tab. 1.

From the results in Tab. 1, we can see that using the Monodepth2 [1] architecture achieves better performance than using P^{2}Net. This is largely due to that the outdoor environments are full of textures. The well-designed Monodepth2 works well in such kinds of scenes, while the strategies adopted in P^{2}Net are more suitable for indoor scenes. This has been discussed in [4]. Though our method does not improve the performance too much by using Monodepth2 architecture, we can see the effectiveness of our extra structural losses by using the P^{2}Net architecture.

Train	RMS↓	AbsRel↓	SqRel↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Using the Monodepth2 architecture						
Original	4.863	**0.115**	0.903	**87.7**	**95.9**	**98.1**
Original-finetune	4.882	0.117	**0.894**	87.2	95.8	98.0
Ours	**4.850**	0.120	0.906	87.0	95.8	**98.1**
Using the P^{2}Net architecture						
Original	5.008	0.121	0.964	86.6	95.4	**97.9**
Original-finetune	5.041	0.121	0.996	**86.7**	95.4	97.8
Ours	**4.969**	**0.120**	**0.941**	**86.7**	**95.5**	**97.9**

Table 1: Outdoor tests using different network architectures on KITTI dataset.

The depth, the surface normal, and the plane detection results of our method on KITTI dataset are shown in Fig. 1. Note that the detected planar regions are mostly located on the road, where the textures are rich enough to supervise a good depth. This may be the major reason why our extra losses did not help too much within the Monodepth2 training pipeline. The other reason may be that the extracted dominant directions may not be strictly mutually perpendicular in outdoor scenes, leading to large surface normal errors.

3. Plane quality tests

We evaluate the plane quality on the IBims-1[2] as [3]. All models are trained on the NYUv2 dataset with the same number of epochs for fair comparison (pretrain epochs are also included in our method). From the results in Tab. 2, as expected, our method produces the best plane quality (the second column), even though P2Net also adopts co-planar losses. The improvements are largely due to the global constraint from Manhattan normal loss. However, all methods produce low structure quality especially the depth edge comparing with supervised methods, indicating great efforts are still required to improve self-supervised depth learning in indoor scenes.

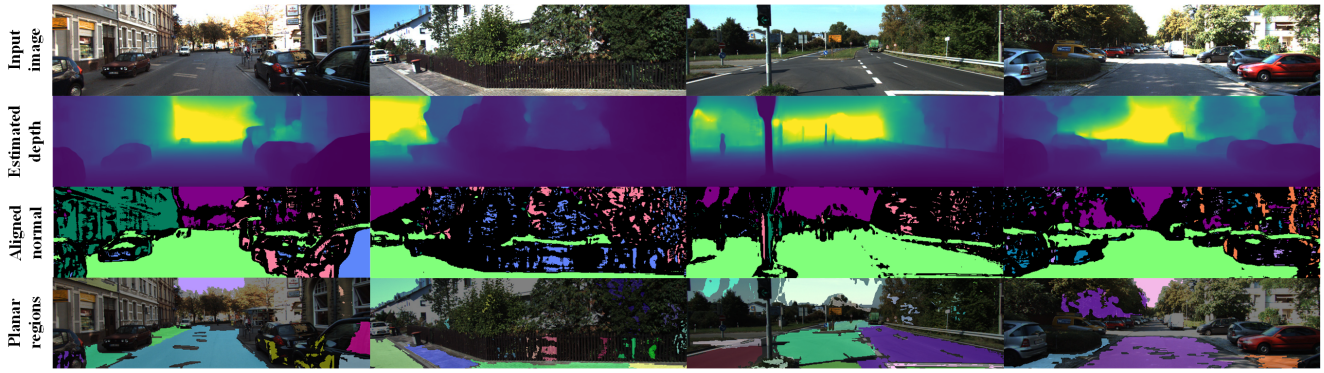

Figure 1: Visualization of the KITTI results. From top to bottom rows: the input image, the estimated depth, the aligned surface normal, and the planar regions detected by our method based on the color and geometric information.

Method	Sup.	$\varepsilon_{\mathrm{DBE}}^{\mathrm{acc}}\downarrow$	$\varepsilon_{\mathrm{DBE}}^{\mathrm{comp}}\downarrow$	$\varepsilon_{\mathrm{PE}}^{\mathrm{plan}}\downarrow$	$\varepsilon_{\mathrm{PE}}^{\mathrm{orie}}\downarrow$	AbsRel$\downarrow$
Wei Yin, et al.[3]	$\surd$	1.90	5.73	2.0	7.41	0.079
Monodepth2[1]	$\times$	**4.455**	68.127	12.160	30.924	**0.220**
P^2Net[4]	$\times$	4.922	67.833	10.823	28.783	0.241
P^2Net-finetune	$\times$	4.628	**49.926**	10.322	28.750	0.232
Ours	$\times$	4.611	67.828	**9.669**	**27.215**	0.227

Table 2: IBims-1 results with the trained model on NYUv2.

References

[1] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[2] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[3] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021.

[4] Zehao Yu, Lei Jin, and Shenghua Gao. P^2net: Patch-match and plane-regularization for unsupervised indoor depth estimation. In *Proceedings of the European conference on computer vision*, 2020.

Figure 2: Point cloud visualization on NYUv2, ScanNet and InteriorNet results. We present the point cloud results of Monodepth2[1], P^{2}Net[4], our method, and the ground-truth. We draw the dominant directions in the scene for better comparison. The results show that our method produces more accurate 3D structures.

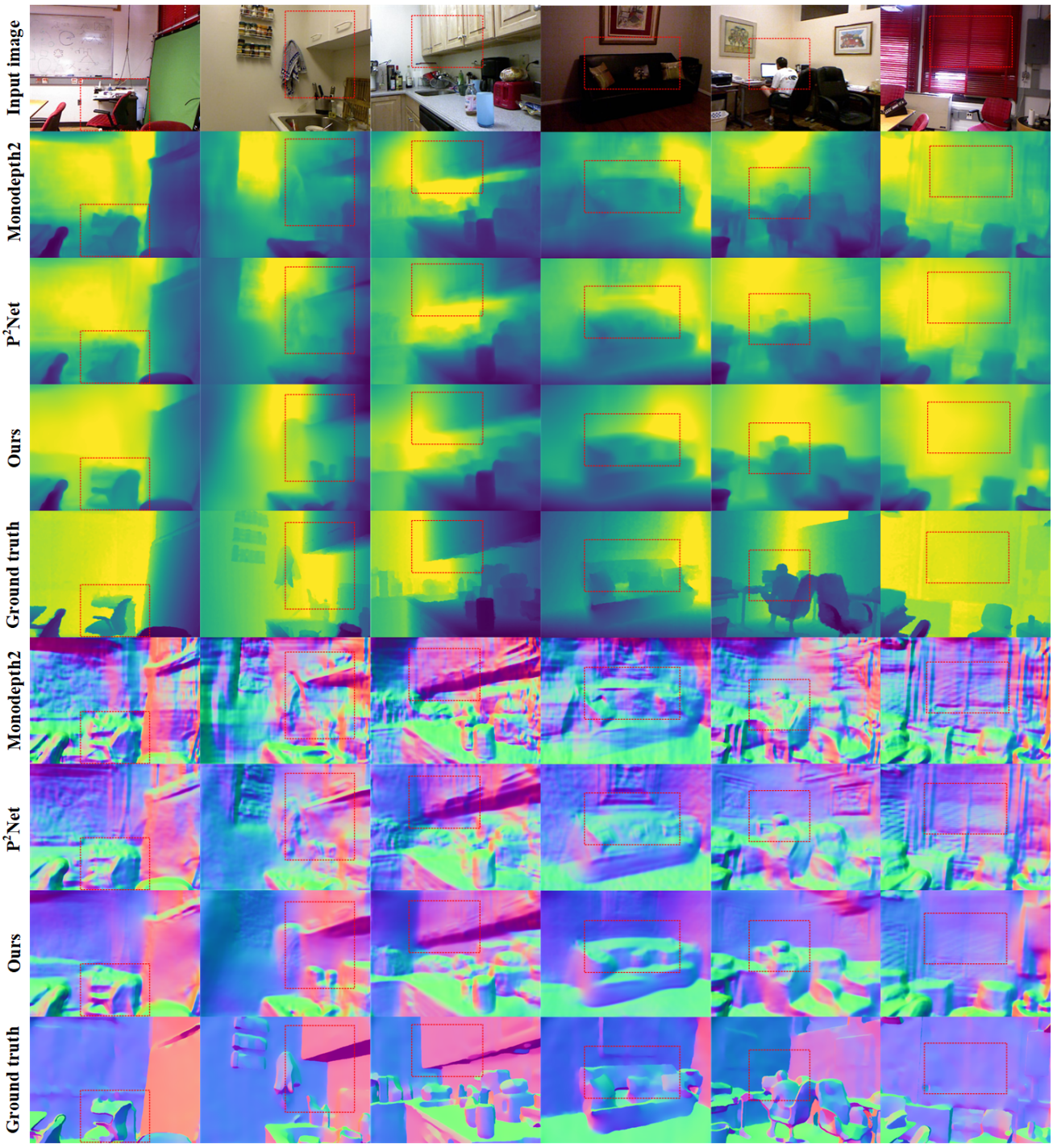

Figure 3: Qualitative visualization results on NYUv2. The top rows show the depth results and the bottom rows show the surface normal results. The results of Monodepth2[1], $\mathrm{P}^2\mathrm{Net}$[4], our method, and the ground-truth depth / normal are presented for comparison. Compared with $\mathrm{P}^2\mathrm{Net}$[4] and Monodepth2[1], our method obtains better surface normal estimation and depth prediction as indicated by the red rectangles.

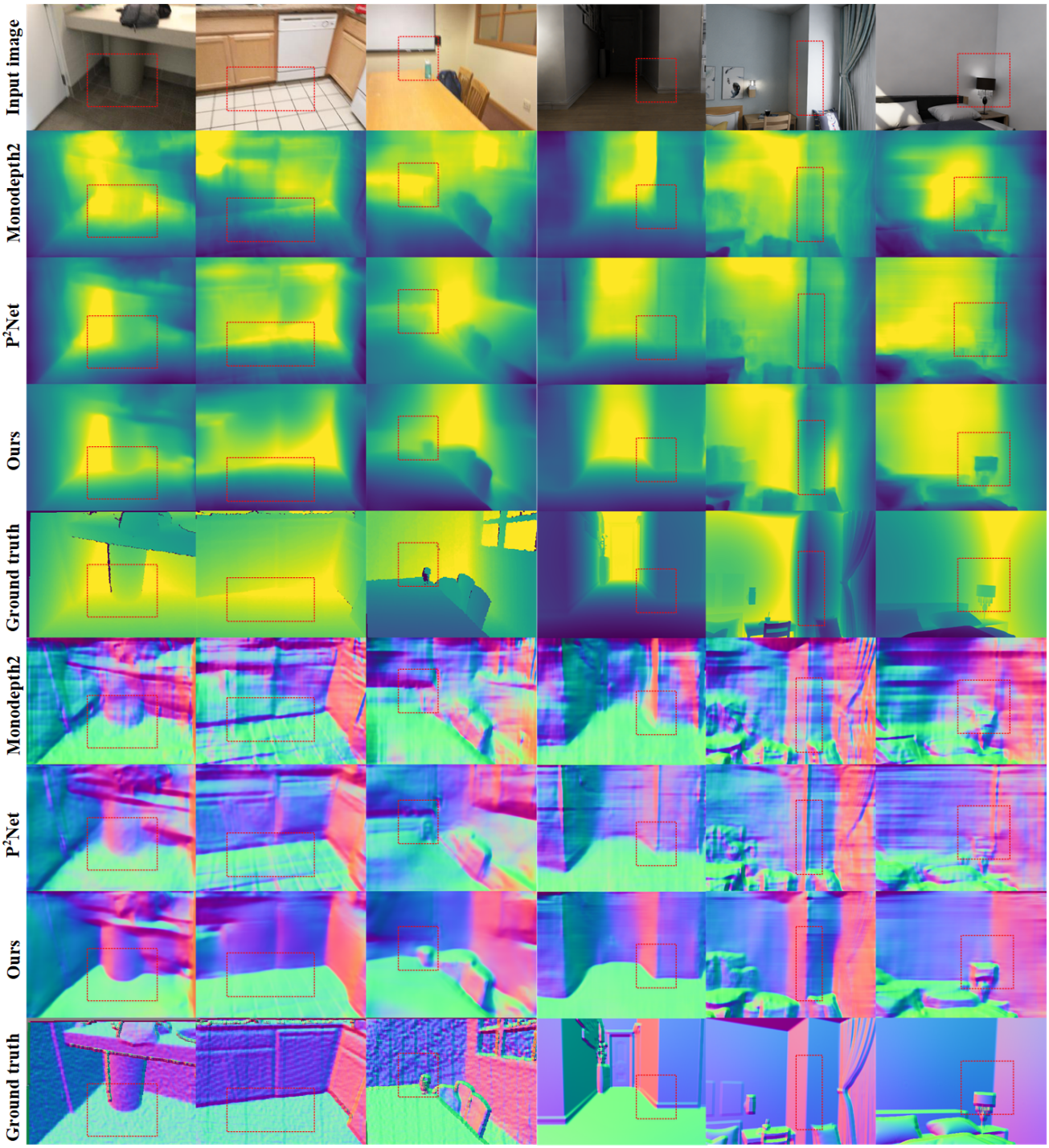

Figure 4: Qualitative visualization results on ScanNet and InteriorNet datasets. The top rows show the depth results and the bottom rows show the surface normal results. The results of Monodepth2[1], $\mathrm{P^2Net}$[4], ours and the ground-truth depth / normal are presented for comparison. Compared with $\mathrm{P^2Net}$[4] and Monodepth2[1], our method obtains better surface normal estimation and depth prediction as indicated by the red rectangles. Our models were trained on the NYUv2 dataset.