Supplementary Materials for
Single-Image Piece-wise Planar 3D Reconstruction via Associative Embedding

Zehao Yu[1*] Jia Zheng[1*] Dongze Lian[1] Zihan Zhou[2] Shenghua Gao[1†]
[1]ShanghaiTech University [2]The Pennsylvania State University
{yuzh,zhengjia,liandz,goashh}@shanghaitech.edu.cn zzhou@ist.psu.edu

In the supplementary materials, we first present the details of our network architecture. We then show some ablation studies of our proposed method. Finally we report additional quantitative and qualitative results on two public datasets: ScanNet [1] and NYUv2 [6].

1. Architecture

Our encoder is an extended version of ResNet-101-FPN [3]. We add two lateral connections and top-down pathways to the original FPN, and the size of resulting feature map from the encoder is $64 \times 192 \times 256$. Three decoders, *i.e.*, plane segmentation decoder, plane embedding decoder, and plane parameter decoder, all share this feature map. Each decoder simply contains a 1×1 convolutional layer. The architecture is shown in Table 1.

Table 1: Network architecture.

Stage	Type	Output Size
Input		$3 \times 192 \times 256$
Encoder	Extended ResNet-101-FPN	$64 \times 192 \times 256$
Plane segm. decoder	1×1 Conv	$1 \times 192 \times 256$
Plane embed. decoder	1×1 Conv	$2 \times 192 \times 256$
Plane param. decoder	1×1 Conv	$3 \times 192 \times 256$

2. Ablation Studies

In this section, we run a number of ablation studies to validate our method. We use plane recall and pixel recall at 0.05m and 0.6m to evaluate the performance of our methods on the ScanNet test set.

Plane parameter. To evaluate the effectiveness of our plane parameter supervisions, we remove either pixel-level parameter supervision L_{PP} or instance-level parameter supervision L_{IP} in this experiment. As shown in Table 2, both terms play an important role in estimating the scene geometry. Figure 1 further visualizes the reconstruction results derived from the predicted pixel-level parameters. We make the following observations: i) the network with pixel-level parameter supervision L_{PP} only produces inconsistent parameters across the entire plane; ii) the network with instance-level parameter supervision L_{IP} only generates reasonably good results w.r.t. the whole scene geometry, but fails produce accurate predictions at pixel level (*e.g.*, the boundary of each plane); iii) with both supervisions, the results are more consistent and stable.

Table 2: Ablation study of plane parameter supervisions on the ScanNet test set. The ✓ indicates the enabled supervision.

Supervision		Per-plane recall		Per-pixel recall	
L_{PP}	L_{IP}	@0.05	@0.60	@0.05	@0.60
✓		20.18	61.16	24.82	75.10
	✓	10.78	62.04	15.72	76.61
✓	✓	**22.93**	**62.93**	**30.59**	**77.86**

Clustering. To validate the efficiency of our mean shift clustering algorithm, we compare our algorithm with vanilla mean shift algorithm in *scikit-learn* [5]. We further analyze the effect of two hyper-parameters: i) the number of anchors per dimension k, ii) the number of iteration T in testing. Experimental results are shown in Table 3. All timings are recorded on the same computing platform with a 2.2GHz 20-core Xeon E5-2630 CPU and a single NVIDIA TITAN Xp GPU. Our proposed method is more efficient, achieving 30 fps on a single GPU. Further, our proposed method is robust to hyper-parameter selection.

3. More Results

In this section, we show more results on the ScanNet and NYUv2 datasets.

Statistics on the number of detected planes. We show some statistics on the number of planes in Figure 3. The histogram illustrates the number of images versus the number of planes. We make the following observations: i) Due

*Equal contribution
†Corresponding author

Table 3: Ablation study of clustering on the ScanNet test set. The * indicates CPU time (with 20 cores). Our method is more efficient and is robust to hyper-parameters selection.

Variant	Hyper-param. k	Hyper-param. T	Per-plane recall @0.05	Per-plane recall @0.60	Per-pixel recall @0.05	Per-pixel recall @0.60	Speed (FPS)
scikit-learn	-	-	22.85	63.13	30.18	76.09	2.86*
Ours	10	10	22.96	62.89	30.64	77.70	32.26
	20	10	22.97	62.96	30.62	**77.80**	22.19
	50	10	23.05	63.11	30.71	77.73	6.69
	10	5	**23.28**	63.65	**30.77**	77.70	**36.10**
	20	5	23.18	**63.72**	30.68	77.58	24.39
	50	5	22.94	63.35	30.41	76.85	8.08

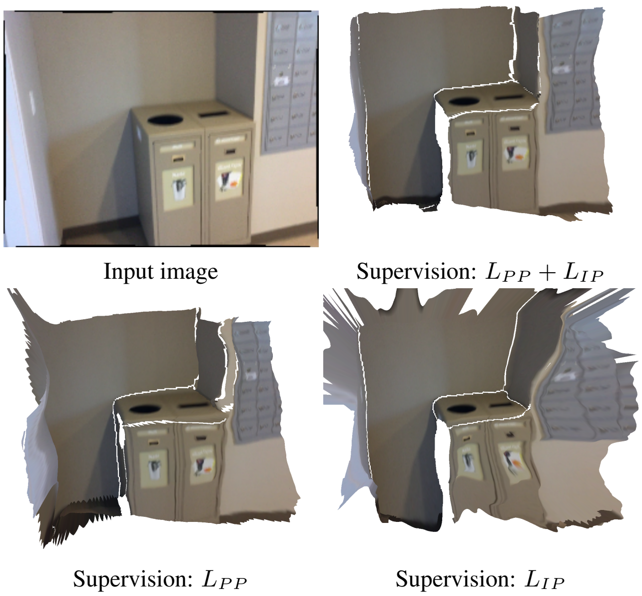


Figure 1: Visualization about plane parameter supervision. Note that all results are reconstructed with the depth maps inferred from pixel-level plane parameters. The results with both supervisions are more consistent and stable.

to the limitation of a fixed number of planes (*i.e.*, 10 planes in PlaneNet), PlaneNet [4] cannot detect all the planes if there are more than 10 planes in the image. ii) Our method is more consistent with the ground truth than PlaneNet.

Quantitative evaluation. We further provide the experiment of depth prediction without fine-tuning on the NYUv2 dataset in Table 4. The results show our method generalizes well.

Besides using depth as threshold, we also use surface normal difference (in degrees) between the predicted plane and ground truth plane as threshold. The threshold varies from 0° to 30° with an increment of 2.5°. As shown in Figure 2, the results are consistent with the results when depth is adopted as threshold. We list the exact numbers of each recall curve in Table 5.

Table 4: Comparison of depth prediction accuracy without fine-tuning on NYUv2 test set. Note that lower is better for top five rows, whereas higher is better for the bottom three rows.

Method	PlaneNet [4]	Ours
Rel	0.238	**0.219**
Rel(sqr)	0.287	**0.250**
$\log_{10}$	0.126	**0.112**
RMSE_{iin}	0.925	**0.881**
$\text{RMSE}_{\log}$	0.334	**0.305**
1.25	49.1	**53.3**
1.25^2	79.0	**84.5**
1.25^3	91.9	**95.1**

Qualitative evaluation. Additional reconstruction results on the ScanNet dataset are shown in Figure 4. More qualitative comparisons against existing methods for plane instance segmentation on the NYUv2 dataset are shown in Figure 5.

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 1

[2] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Manhattan-world stereo. In *CVPR*, pages 1422–1429, 2009. 6

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 1

[4] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *CVPR*, pages 2579–2588, 2018. 2, 5, 6

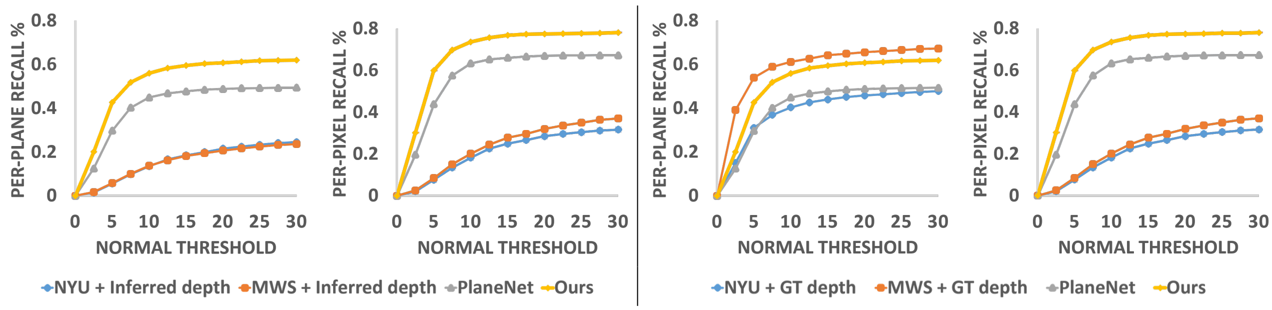


Figure 2: Plane and pixel recall curves with normal difference as threshold on the ScanNet dataset. Our method obtains consistent results when surface normal difference is adopted as threshold.

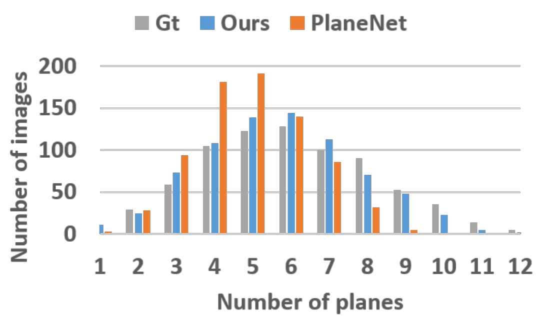


Figure 3: The number of images versus the number of planes in the image.

[5] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *JMLR*, 12:2825–2830, 2011. 1

[6] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012. 1, 5, 6

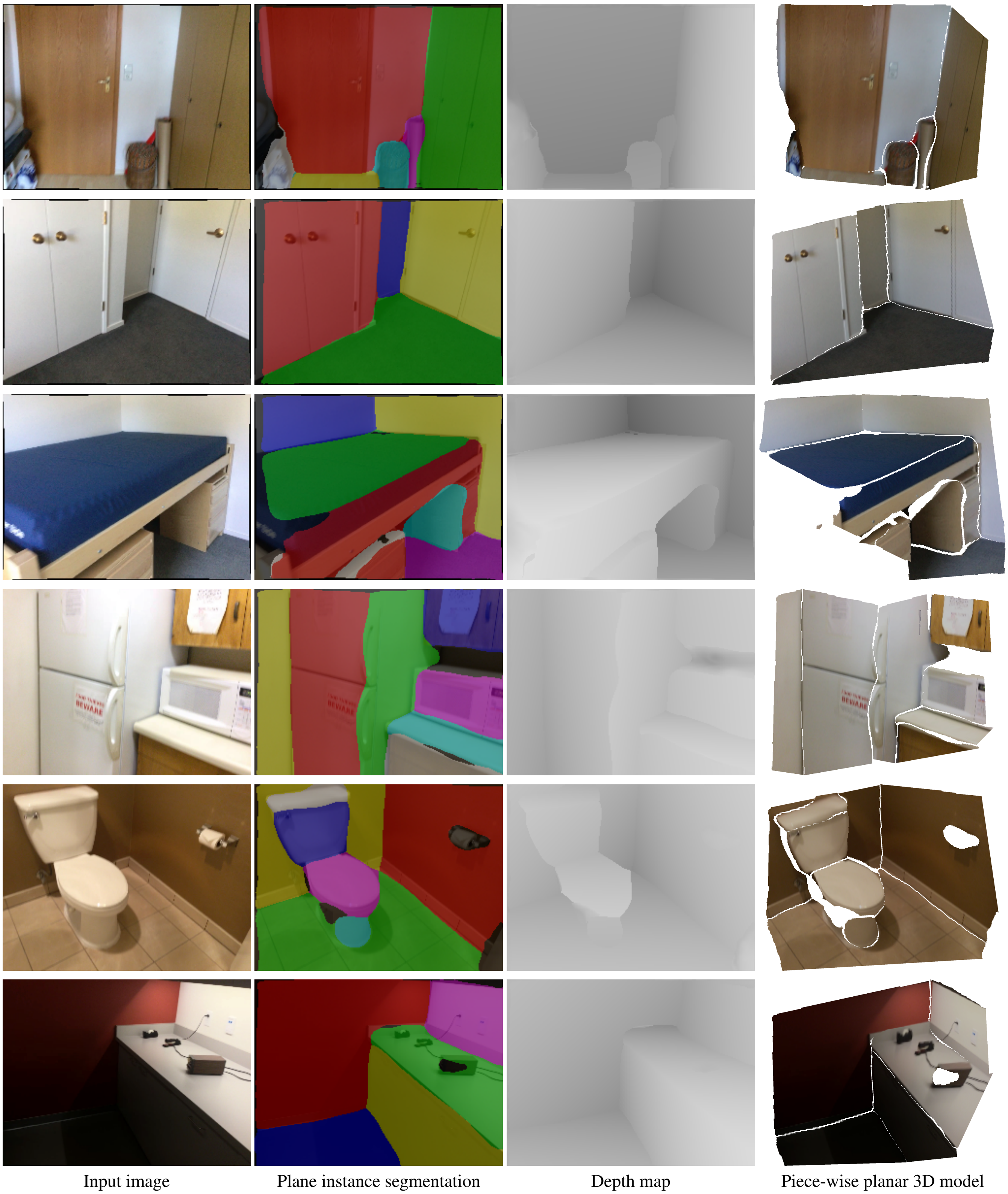
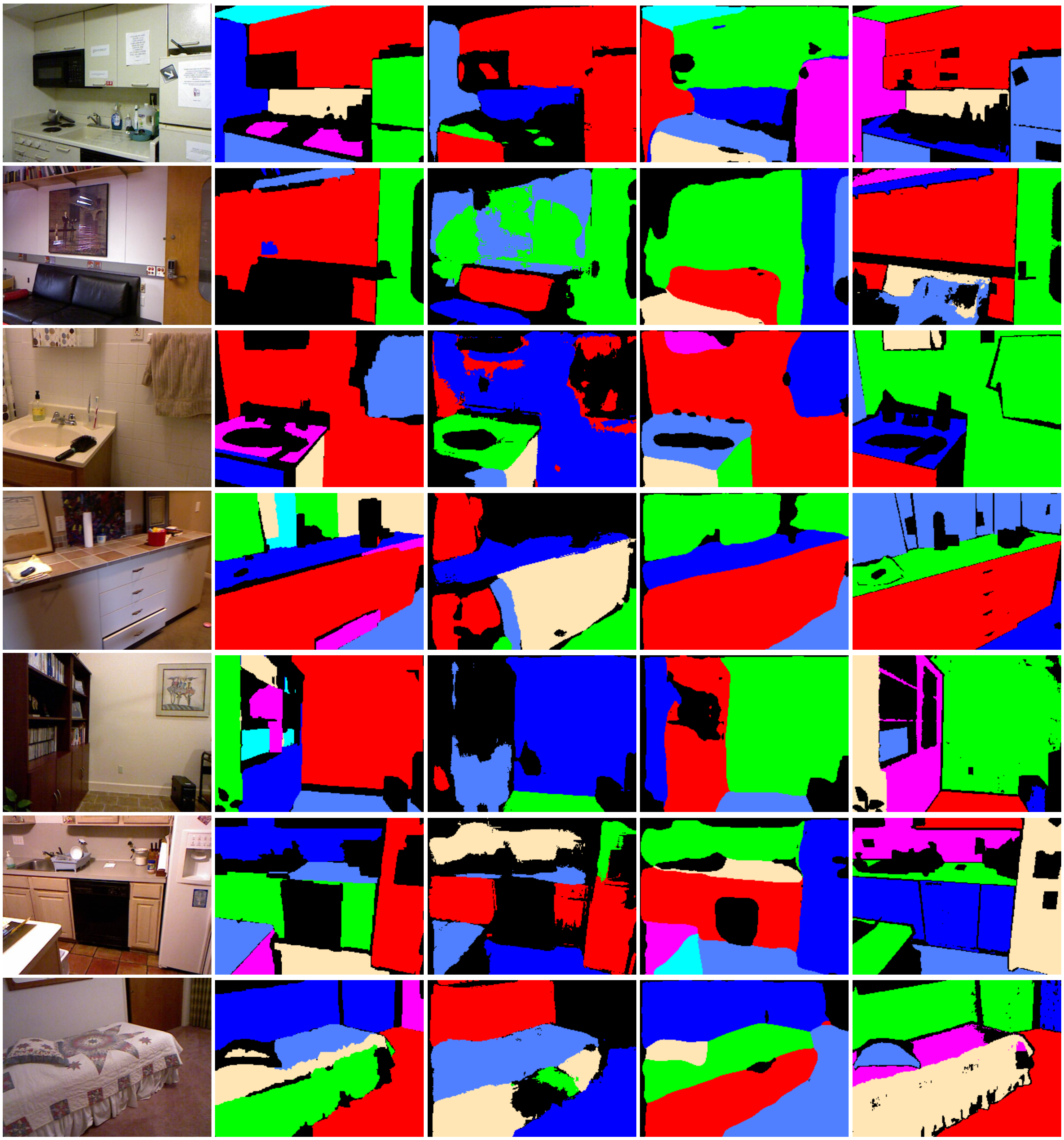


Figure 4: More piece-wise planar reconstruction results on the ScanNet dataset. In the plane instance segmentation results, black color indicates non-planar regions.



Input image GT Depth + [6] PlaneNet [4] Ours Ground truth

Figure 5: More plane instance segmentation results on the NYUv2 dataset. Black color indicates non-planar regions.

Table 5: Plane reconstruction accuracy comparisons on the ScanNet dataset.

(a) Plane recall versus depth difference.

Depth threshold		0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60
GT Depth	MWS [2]	51.22	63.84	67.20	68.28	68.61	68.74	68.85	68.87	68.89	68.92	68.92	68.92
	NYU-Toolbox [6]	45.66	48.34	48.69	48.82	48.89	48.91	48.91	48.93	48.93	48.93	48.96	48.96
Inferred Depth	MWS [2]	1.69	5.32	8.84	11.67	14.40	16.97	18.71	20.47	21.68	23.06	24.09	25.13
	NYU-Toolbox [6]	3.14	9.21	13.26	16.93	19.63	21.41	22.69	23.48	24.18	25.04	25.50	25.85
	PlaneNet [4]	15.78	29.15	37.48	42.34	45.09	46.91	47.77	48.54	49.02	49.33	49.53	49.59
	Ours	22.93	40.17	49.40	54.58	57.75	59.72	60.92	61.84	62.23	62.56	62.76	62.93

(b) Pixel recall versus depth difference.

Depth threshold		0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60
GT Depth	MWS [2]	64.44	74.37	76.36	76.85	76.96	77.03	77.07	77.08	77.09	77.09	77.09	77.09
	NYU-Toolbox [6]	73.59	75.49	75.67	75.75	75.78	75.80	75.80	75.80	75.80	75.80	75.81	75.81
Inferred Depth	MWS [2]	2.40	8.02	13.70	18.06	22.42	26.22	28.65	31.13	32.99	35.14	36.82	38.09
	NYU-Toolbox [6]	3.97	11.56	16.66	21.33	24.54	26.82	28.53	29.45	30.36	31.46	31.96	32.34
	PlaneNet [4]	22.79	42.19	52.71	58.92	62.29	64.31	65.20	66.10	66.71	66.96	67.11	67.14
	Ours	30.59	51.88	62.83	68.54	72.13	74.28	75.38	76.57	77.08	77.35	77.54	77.86

(c) Plane recall versus normal difference.

Normal threshold		2.5	5.0	7.5	10.0	12.5	15.0	17.5	20.0	22.5	25.0	27.5	30.0
GT Normal	MWS [2]	39.19	54.03	58.93	61.23	62.69	64.22	64.90	65.58	66.15	66.61	67.13	67.29
	NYU-Toolbox [6]	15.04	31.07	37.00	40.43	42.66	44.02	45.13	45.81	46.36	46.91	47.41	47.82
Inferred Normal	MWS [2]	1.73	05.79	10.04	13.71	16.23	18.22	19.48	20.71	21.69	22.50	23.25	23.60
	NYU-Toolbox [6]	1.51	05.58	09.86	13.47	16.64	18.48	19.99	21.52	22.48	23.33	24.12	24.54
	PlaneNet [4]	12.49	29.70	40.21	44.92	46.77	47.71	48.44	48.83	49.09	49.20	49.31	49.38
	Ours	20.05	42.66	51.85	55.92	58.34	59.52	60.35	60.75	61.23	61.64	61.84	61.93

(d) Pixel recall versus normal difference.

Normal threshold		2.5	5.0	7.5	10.0	12.5	15.0	17.5	20.0	22.5	25.0	27.5	30.0
GT Normal	MWS [2]	56.21	70.53	73.49	74.47	75.12	75.66	75.88	76.04	76.28	76.41	76.55	76.59
	NYU-Toolbox [6]	31.93	58.92	65.63	69.09	71.12	72.10	72.89	73.41	73.65	74.08	74.39	74.65
Inferred Normal	MWS [2]	2.58	8.51	15.08	20.16	24.51	27.78	29.63	31.96	33.65	34.99	36.37	37.03
	NYU-Toolbox [6]	2.11	7.69	13.49	18.25	22.58	24.92	26.63	28.50	29.58	30.46	31.23	31.65
	PlaneNet [4]	19.68	43.78	57.55	63.36	65.27	66.03	66.64	66.99	67.16	67.20	67.26	67.29
	Ours	30.20	59.89	69.79	73.59	75.67	76.8	77.3	77.42	77.57	77.76	77.85	78.03