PlaneRecTR: Unified Query Learning for 3D Plane Recovery from a Single View

Jingjia Shi   Shuaifeng Zhi*   Kai Xu*
National University of Defense Technology, China

Appendix

A. Supplementary Video

We encourage readers to also watch our *supplementary video* (https://youtu.be/YBB7totHGJg) which gives a more vivid illustration of our pipeline and shows 3D reconstruction results.

B. Details of Training and Inference

In the ScanNet experiments, we screened out 2307 images from the complete training set (50,000 images) with poor plane parameter annotations, while the testing set remained the same as the previous methods to fair comparison. During inference, we dropped non-plane predictions from a total of N predicted plane probabilities. Sigmoid function is then performed on the remaining K plane predictions to obtain plane-level soft masks, followed by an argmax operation along K soft masks to obtain plane segmentation. During this process, if the peak soft mask value of certain pixels is below a pre-defined threshold (*i.e.*, 0.001 in our implementation), they are also regarded as non-plane regions. We find this could prevent the network from excess plane predictions on non-planar structures.

C. More Evaluation Metrics on NYUv2-Plane

As our method only estimates the depth of planar regions, for a fair comparison we also conduct evaluations using the same reconstruction metrics to ScanNet: per-pixel/plane recall rates of depth, normal, respectively. Based on similar definitions in section 4 ”Evaluation Metrics” of the main text, per-pixel/plane recall of plane offset is added and its threshold varies from 25mm to 300mm with an increment of 25mm (see Figure 1 and Table 1 for details). Similarly, our method variants with all three backbones overall outperform PlaneTR [3] and we see a large performance boost of PlaneRecTR (HRNet-32 [4]/Swin-B [2]) against the ResNet-50 [1] counterpart. In addition, we show the average error statistics of plane parameters separately in Table 2, we observe obvious accuracy improvement using HRNet-32/Swin-B backbone, especially on the offset estimation task on unseen NYUv2-Plane data.

Method	Per-Pixel/Per-Plane Recalls ↑					
	Depth		Normal		Offset	
	@0.10 m	@0.60 m	@5°	@30°	@50 mm	@300 mm
PlaneTR [3]	7.08/5.07	41.98/27.10	20.08/11.69	52.08/32.85	9.06/5.71	35.64/22.75
PlaneRecTR	**7.72/6.48**	**44.44/35.70**	14.43/10.56	**55.99/42.24**	**10.30/6.97**	**38.51/28.69**
PlaneRecTR (HRNet-32)	**8.99/8.06**	**48.60/39.33**	19.91/**14.06**	**58.58/45.27**	10.04/6.81	**41.52/31.55**
PlaneRecTR (Swin-B)	**10.58/9.36**	**54.06/42.40**	**24.08/16.42**	**59.92/45.68**	8.94/6.41	**44.30/33.10**

Table 1. Per-pixel and per-plane recalls comparison on the NYUv2-Plane dataset.

Method	Plane Parameter Estimation Errors ↓	
	Normal (°)	Offset (mm)
PlaneTR [3]	17.09	615.92
PlaneRecTR	**15.98**(-1.11)	**611.82**(-4.10)
PlaneRecTR (HRNet-32)	**15.55**(-1.54)	**577.28**(-38.64)
PlaneRecTR (Swin-B)	**15.08**(-2.01)	**553.47**(-62.45)

Table 2. Plane parameters estimation comparison on the NYUv2-Plane dataset.

D. More Details of Figure 5

Since the performance of PlaneRecTR and PlaneRecTR (HRNet-32) in Figure 5 of the main text overlaps, we show the specific values of some methods under lower and higher thresholds respectively in Table 3.

Method	Per-Pixel/Per-Plane Recalls ↑			
	Depth		Normal	
	@0.10 m	@0.60 m	@5°	@30°
PlaneTR [3]	52.89/40.76	80.52/61.49	59.45/43.14	80.25/60.68
PlaneRecTR	**53.07/45.07**	**83.60/72.84**	**62.75/48.48**	**83.85/71.33**
PlaneRecTR (HRNet-32)	**54.32/47.36**	**83.73/73.67**	**62.79/49.07**	83.81/**71.92**
PlaneRecTR (Swin-B)	**57.74/50.03**	**85.34/75.49**	**67.43/52.02**	**85.19/73.67**

Table 3. Per-pixel and per-plane recalls comparison on the ScanNet dataset.

[1] Shuaifeng Zhi and Kai Xu are corresponding authors.

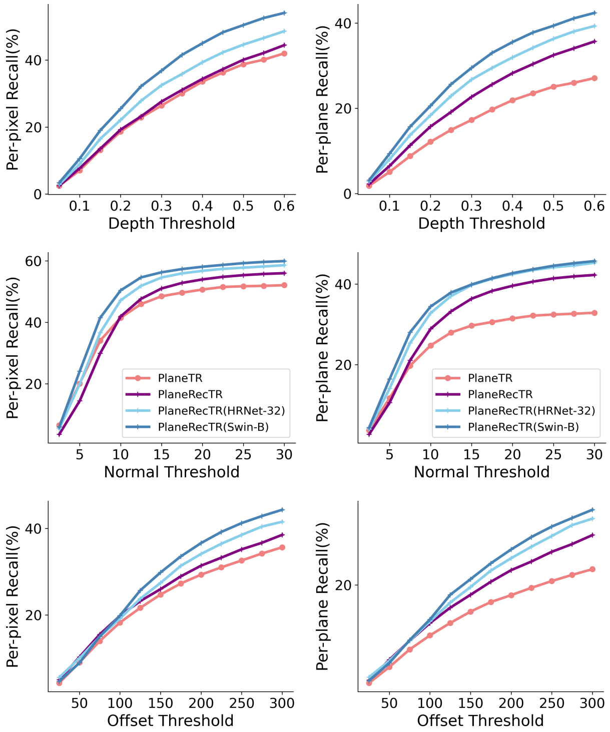


Figure 1. Per-pixel and per-plane recalls on the NYUv2-plane dataset.

E. A Visual Comparison of PlaneRecTR and PlaneRecTR (Swin-B)

In Figure 2 we show qualitative results of our proposed methods in details. When using a powerful backbone model, the network is able to predict more accurate plane segmentation masks (row 2,7,8), discriminate confusing nearby planes (row 3,8), and avoid over segmentation (row 4). In it also exciting to see that PlaneRecTR (Swin-B) even discover small planes which are missed by PlaneRecTR (row 1,5,6) and the ground truth (row 1,5).

F. Supplementary Visualization

In Figure 3 and Figure 4, we add 3D segmentation masks visualization for Figure 3 of the main text and comparison of depth and 3D segmentation for Figure 4 of the main text in order to better display 3D plane recovery results.

G. More Visualization Results of PlaneRecTR

We display more visual comparisons of our method on both datasets in Figure 5.

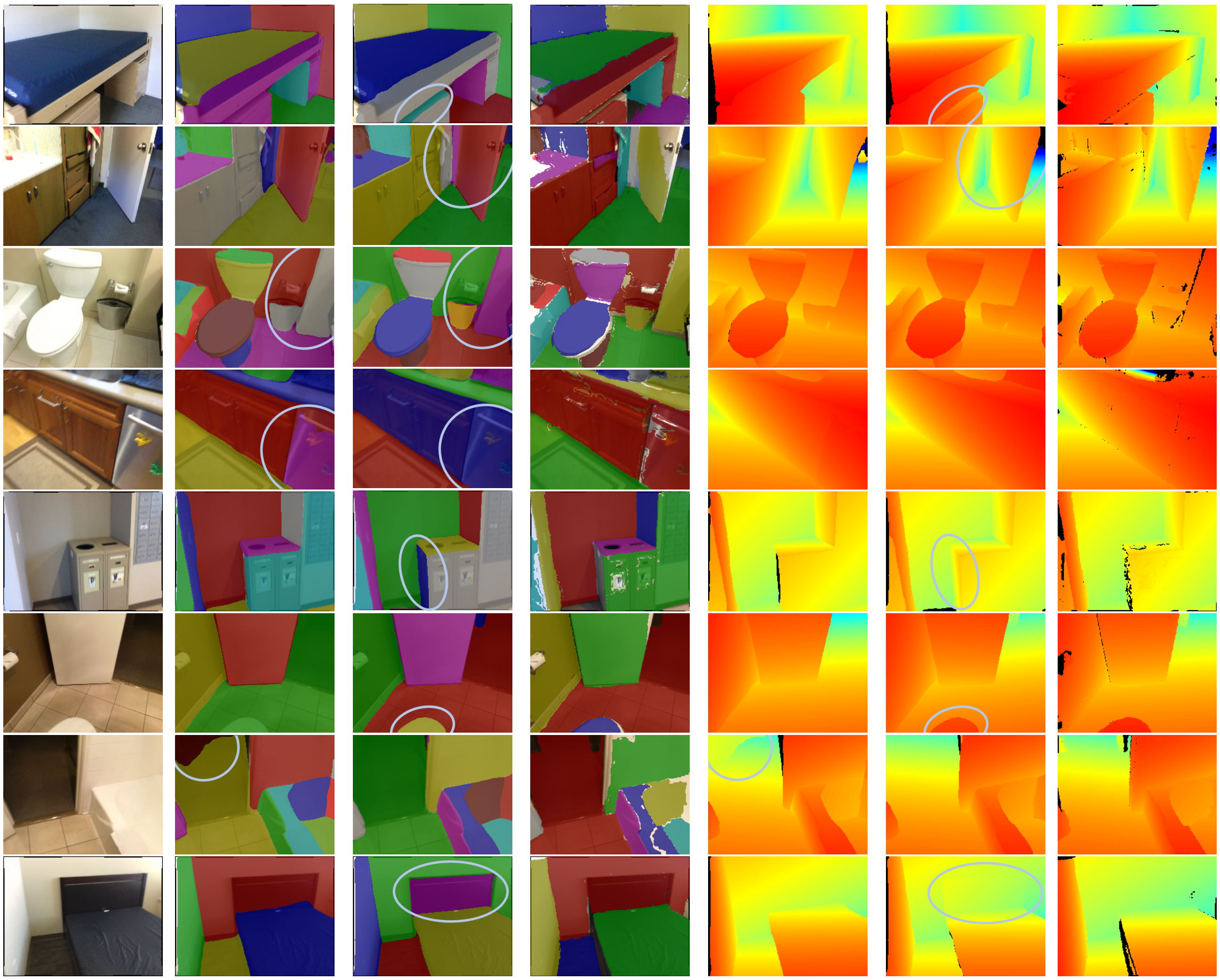
References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1

[3] Bin Tan, Nan Xue, Song Bai, Tianfu Wu, and Gui-Song Xia. Planetr: Structure-guided transformers for 3d plane recovery. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1

[4] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *CoRR*, abs/1908.07919, 2019. 1



(a) Input (b) ResNet50 Mask (c) Swin-B Mask (d) GT Mask (e) ResNet50 Depth (f) Swin-B Depth (g) GT Depth

Figure 2. Comparison of plane reconstruction results of different backbones on the ScanNet dataset.

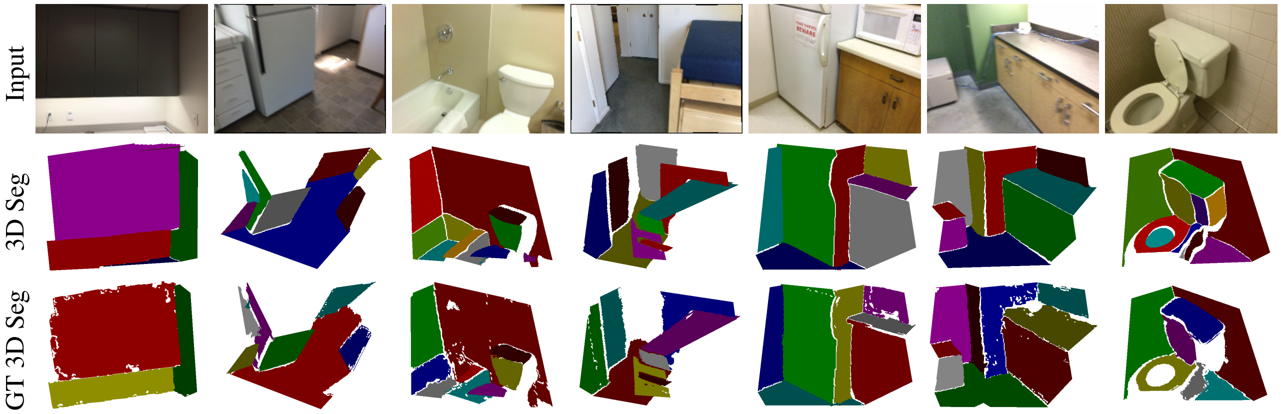


Figure 3. 3D segmentation masks of PlaneRecTR on the ScanNet dataset.

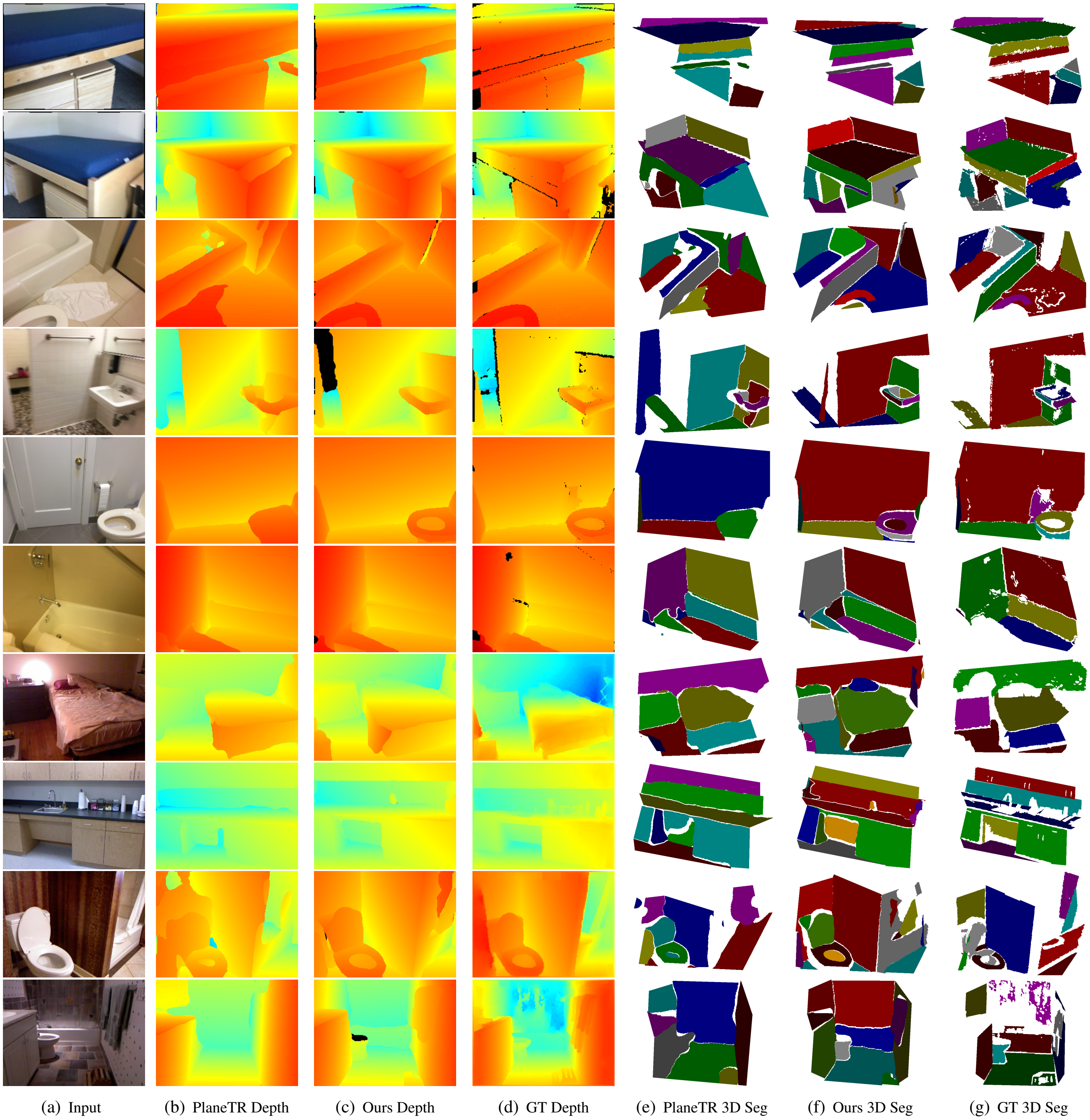


Figure 4. Comparison of 3D segmentation masks on the ScanNet (top 6 rows) and NYUv2-Plane (bottom 4 rows) datasets.

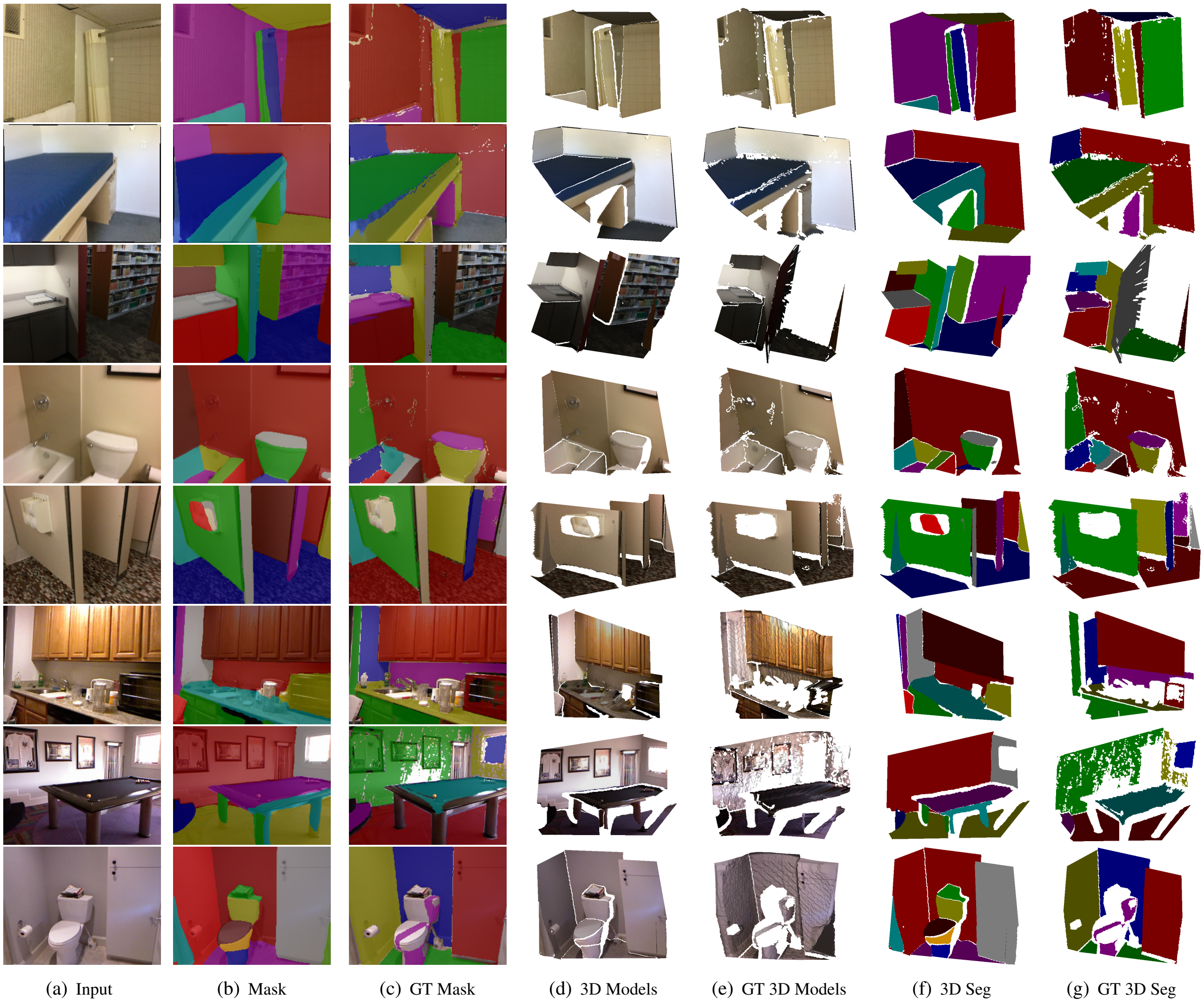


Figure 5. More 3D plane recovery results on the ScanNet (top 5 rows) and NYUv2-Plane (bottom 3 rows) datasets.