PanoContext-Former: Panoramic Total Scene Understanding with a Transformer

Supplementary Material

This supplementary material contains:

- Quantitative evaluation and more visualized samples of **ReplicaPano** dataset.
- Implementation details of both our method and compared method, including training strategies, network architecture, and parameter settings.
- More quantitative comparisons of 3D detection on all categories of iGibson-Synthetic [16] and ReplicaPano.
- More quantitative comparisons with other approaches on 3D detection and room layout estimation.
- More qualitative results on iGibson-Synthetic and ReplicaPano.

.

1. ReplicaPano Dataset Samples

ReplicaPano is a new panoramic dataset that offers various ground truths, including photo-realistic panorama, depth maps, real-world 3D room layouts and 3D oriented object bounding boxes, and object meshes (Fig. 1). The RGB panoramic images and depth maps are all rendered from real-scan [11]. As shown in Fig. 2, we hired three data annotators and spent over 700 hours annotating the room layout and oriented object bounding boxes through PanoAnnotator [15] and labelCloud [9], respectively. To get a consistent room layout among all panoramas of the same scene, the annotation process for the room layout involves two steps: First, we select several panoramas in different positions of the same room and annotate each of them. We then fuse and refine these labeled layouts to obtain the final room layout. To ensure high-quality annotations, the object bounding boxes of the same scene are annotated by two annotators and verified by a third annotator. The object meshes are selected manually, and then scaled and transformed to match the scene mesh using Blender. We conducted quantitative evaluations for the annotations of ReplicaPano (Tab. 3 and Tab. 4). For room layout, we calculate the 2D-IoU between the floor plan mask from the raw scene mesh and the labeled room layout. For object shape, we calculate the Partial Chamfer Distance between the aligned CAD model and the incomplete object mesh of the scene. In order to facilitate the community, we will release not only the dataset but also the rendering codes.

2. Limitations

Although our method achieves state-of-the-art performance on the panoramic scene understanding tasks, there are still limitations in our model. Our method fails to robustly detect objects with thin structures, such chair, table, and floor-lamp. It could be because these objects have sparse points on the thin structure and may be occluded by other objects, making it difficult for the model to accurately detect them. Another limitation is that there are some objects have inconsistencies between mesh and image contents in the **ReplicaPano** dataset. Additionally, the dataset volume is smaller than other popular 3D scene datasets such as MP3D [1] and ScanNet [3]. We are making effort to expand the real-world data and refine the object meshes to address these problems. For the ONet representation, although the compact representation makes it efficient for inference, it may not be able to recover very fine-grained structure. We believe that increasing the shape code dimension and training it on a more diverse set of object datasets could improve its generalization performance and enable it to handle a wider range of objects.

Loss Weight		Term	
Symbol	Value	Symbol	Description
λ_p	1.0	$\mathcal{L}_{pos}$	Layout positional loss
λ_n	1.0	$\mathcal{L}_{norm}$	Layout orientation loss
λ_e	0.1	$\mathcal{L}_{sharp}$	Layout sharpness loss
β_{samp}	0.2	$\mathcal{L}_{\text{samp}}$	Object BBox sampling loss
β_{objness}	0.4	$\mathcal{L}_{\text{objness}}$	Object BBox objectness loss
β_{cls}	0.1	$\mathcal{L}_{\text{cls}}$	Object BBox classification loss
β_{cen}	0.1	$\mathcal{L}_{\text{cen}}$	Object BBox center offset loss
$\beta_{\text{size\_cls}}$	0.1	$\mathcal{L}_{\text{size\_cls}}$	Object BBox size classification loss
$\beta_{\text{size\_off}}$	0.06	$\mathcal{L}_{\text{size\_off}}$	Object BBox size offset loss
$\beta_{\text{head\_cls}}$	0.1	$\mathcal{L}_{\text{head\_cls}}$	Object BBox head classification loss
$\beta_{\text{head\_off}}$	0.04	$\mathcal{L}_{\text{head\_off}}$	Object BBox head offset loss
β_{shape}	1e-5	$\mathcal{L}_{\text{shape}}$	Object shape code loss
σ_l	1.0	$\mathcal{L}_{layout}$	Layout loss
σ_o	1.0	$\mathcal{L}_{object}$	Object loss
σ_p	0.5	$\mathcal{L}_{physic}$	Physical violation loss

Table 1. Joint training losses and their corresponding weights.

3. Implementation Details

3.1. Training strategy of PanoContext-Former

For monocular depth estimation, Unifuse [4] is finetuned on iGibson-Synthetic and ReplicaPano using the weights pretrained on Matterport3D [1], the batch size is 6 and learning rate is 1e-4 for 100 epochs. For 3D autoencoder, ONet [6] is finetuned from the weights pretrained on ShapeNet [2],

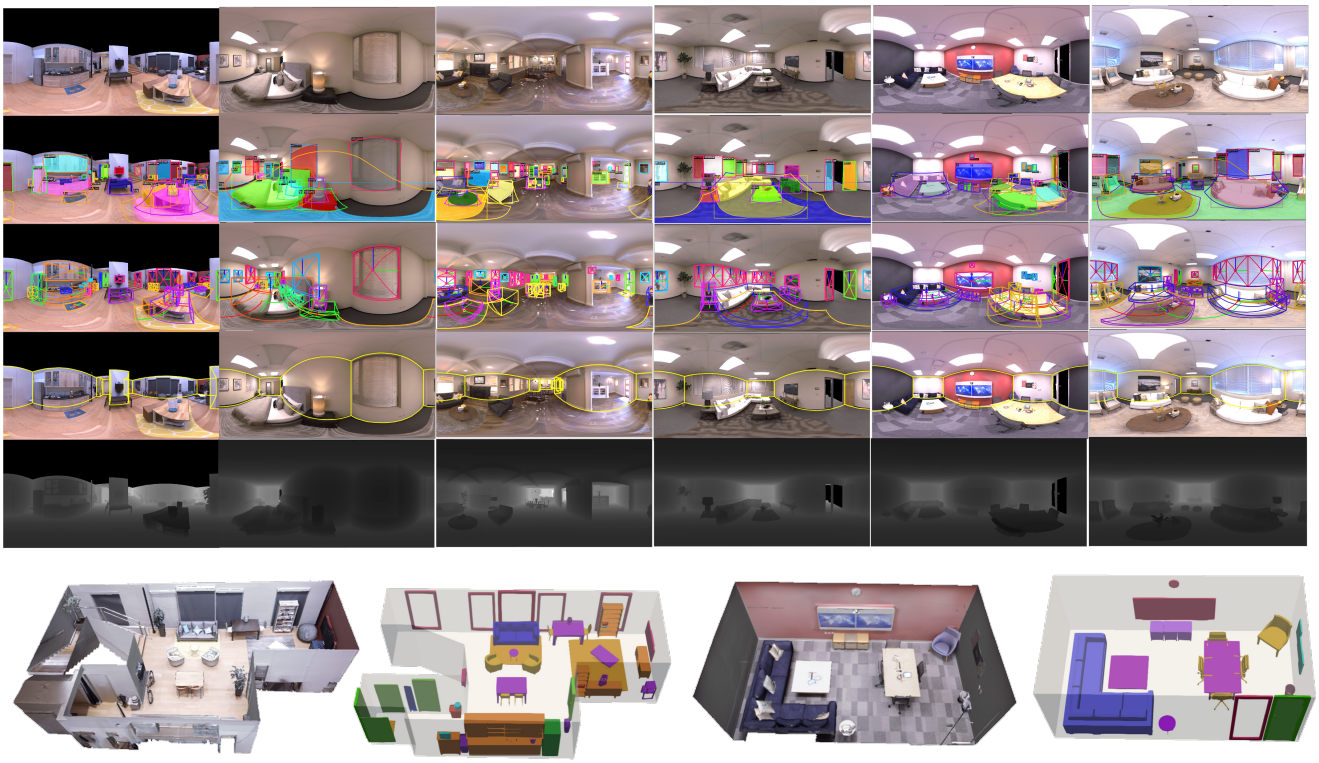

Figure 1. Samples of ReplicaPano. In the top row, there are photo-realistic panoramas rendered from Replica. The second row shows the 2D bounding box of each object. The 3D oriented bounding boxes of objects are shown in the third row. The fourth and fifth rows contain each room's layout and the high-fidelity depth image. The bottom row demonstrates the gravity-aligned scene meshes and the ground truth of full scene reconstruction.

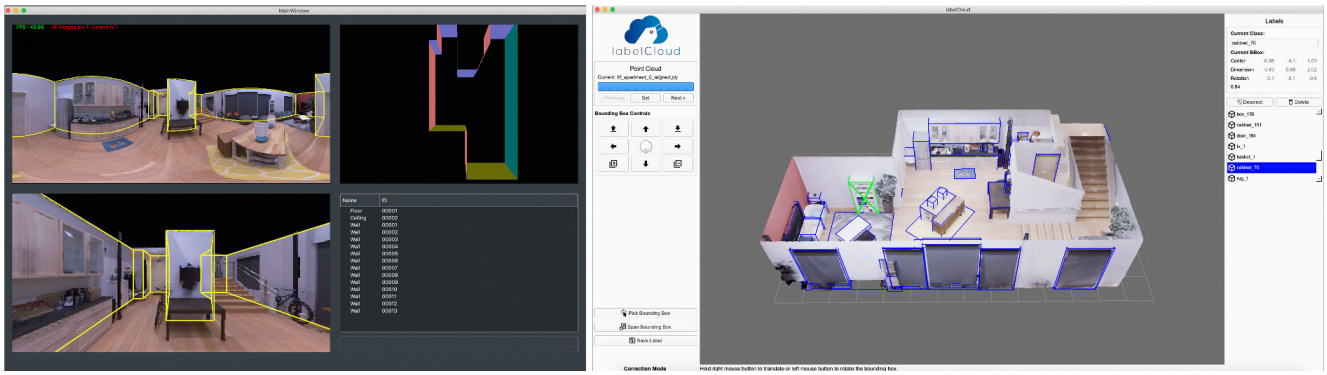

Figure 2. ReplicaPano annotation example. We use Panoanotator [15] for layout annotation (left) and lableCloud [9] for oriented bounding box annotation (right).

Method	Params	Depth Estimation	Object Reconstruction	Other Module	Total
DeepPanoContex	132.08M	-	1.85s	5.06s (ODN,LEN,GCN,RO)	6.91s
Ours	88.35M	0.08s	0.11s	0.32s (ODN,LEN,CM,PP)	0.51s

Table 2. The inference time and parameters comparison on V100.

Corners	4 corners	6 corners	8 corners	10+ corners	odd corners	overall
2D-IoU	98.44	97.40	97.69	96.58	97.22	97.47

Table 3. Quantitative evaluation of the room layout annotation.

with batch size of 64 and learning rate of 1e-4 for 300000 iterations. We train ODN and LEN jointly from scratch

Cat.	cabinet	door	chair	curtain	lamp	rug	sofa	table	trash	tv
CD	0.15	0.028	0.23	0.33	0.039	0.019	0.26	0.15	0.042	0.034

Table 4. Quantitative evaluation of object shape annotation. The Partial Chamfer Distance (from the incomplete object mesh to the CAD model) scaled with the factor of 10^2.

by the AdamW optimizer (β_1 =0.9, β_2 =0.999) with batch size of 4 for 400 epochs. In ODN module, the input is 50K points with data augmentation including random flip and panoramic horizontal rotation. The initial learning rates of ODN, LEN, and context module are set to 0.004, 0.0001, and 0.0002, respectively, followed by the cosine decay schedule. The loss weights and their corresponding descriptions in joint learning are shown in Tab. 1.

3.2. Network Architecture of PanoContext-Former

For ODN module, we follow [5] to use PointNet++ as the backbone network. There are four set abstraction layers and two feature propagation layers in the backbone. For each set abstraction layer, the input point cloud is sub-sampled to 2048, 1024, 512, and 256 points with the increasing receptive radius of 0.2, 0.4, 0.8, and 1.2, respectively. And the two feature propagation layers up-sample the point cloud features to 512 and 1024. Then we generate 256 initial object proposals from point cloud features. For LEN module, we employ ResNet-18 to encode image features from panorama following [7]. In addition, we extracted features from perspective views and fused them with panoramic features by E2P-based feature fusion. The fused features are fed into the first GCN block which is composed of 6 GCN layers followed by a linear layer and returns 642 vertex offsets of a tessellated sphere. As described in our paper, all the above features with their position embeddings are concatenated together and act as input to the transformer-based context module. On each transformer encoder layer, we apply the object head and shape head to estimate the oriented 3D bounding box and shape of each object, and ensemble the result of each layer. The two heads share two layers of multi-layer perception and have an independent linear layer separately. The refined layout features and the associated vertices are subdivided, then fed into the layout head to generate the final layout which has 2562 vertices and 5120 faces. We compare the running time and model complexity with DeepPaonContext [16] in Tab. 2.

3.3. Network Details of the Compared Methods

3D Object Detection: For DeepPanoContext-3D, we first get the point cloud of the object based on the 2D detection and depth estimation results; then the point cloud is downsampled to 1024 and fed into PointNet++ encoding geometry features of dimension 1024; finally, these features are concatenated with appearance features and relational features for object bounding box estimation. For Group-Free [5], we use the official implementation with the 50k points as input and choose the best parameters setting for comparison. For DeepPanoContext, Im3D-Pano, and Total3D-Pano, we use the results reported on [16].

3D Room Layout Estimation: We train HorizonNet [12], HoHoNet [13], Led2-Net [14], Deep3DLayout [7], and DOPNet [10] from the source code provided by each author with the default parameter settings.

4. 3D detection Results on All Categories

In Tab. 7 and Tab. 8, we show all-category results on iGibson-Synthetic and ReplicaPano respectively. Our method outperforms other baselines on most categories and the average mAP, which is consistent with the conclusion drawn in the main paper.

5. Comparison with Other Methods

According to Tab. 5 and Tab. 6, which compare recent works that focus on room layout estimation and 3D object detection, our approach still outperforms these works.

Method	chair	soft	table	fridge	sink	door	floor lamp	bottom cabinet	top cabinet	sofa chair	dryer	mAP↑
TR3D	35.81	90.47	59.2	**83.80**	**95.38**	**92.28**	**53.99**	40.6	73.08	69.86	10.77	64.11
Ours	**38.47**	**98.15**	**66.61**	82.77	89.55	87.49	40.31	**59.53**	**80.71**	**83.42**	**13.83**	**67.35**

Table 5. 3D detection comparison with TR3D [8] on iGibson.

Method	2D-IoU ↑	3D-IoU ↑
DOPNet [10]	90.96	90.63
Ours	**92.24**	**92.04**

Table 6. Layout estimation compared with DOPNet on iGibson.

6. More Qualitative Results

We also show more qualitative results of our method on iGibson-Synthetic (Fig. 3) and ReplicaPano (Fig. 4). The full scene reconstruction results illustrate that our method is capable to generate not only accurate 3D object bounding boxes and poses but also room layout and object meshes with good visual quality.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

Method	chair	sofa	table	fridge	sink	door	floor lamp	bottom cabinet	top cabinet	sofa chair	dryer
Total-Pano	20.84	69.65	31.79	43.13	68.42	10.27	16.42	34.42	20.83	62.38	33.78
Im3D-Pano	33.08	72.15	37.43	70.45	75.20	11.58	6.06	43.28	18.99	78.46	41.02
DeepPanoContext	27.78	73.96	46.85	74.22	75.29	21.43	20.69	52.03	50.39	77.09	59.91
DeepPanoContext-3D	**39.41**	78.03	51.44	75.24	81.46	51.97	**60.01**	55.56	42.58	79.99	**60.07**
Group-Free	27.83	96.04	61.57	**84.69**	87.69	82.20	27.20	56.46	77.99	79.21	8.29
Ours	38.47	**98.15**	**66.61**	82.77	**89.55**	**87.49**	40.31	**59.53**	**80.71**	**83.42**	13.83

Method	window	carpet	picture	oven	bottom cabinet no top	counter	dish washer	shelf	coffee table	toilet	mirror
Total3D-Pano	3.07	0.05	0.02	31.33	34.40	0.78	43.54	10.93	39.72	90.00	0.11
IM3D-Pano	3.42	0.01	0.01	29.06	44.79	1.34	43.80	15.41	56.82	90.00	0.16
DeepPanoContext	9.56	0.65	0.21	34.50	44.17	1.25	63.19	22.65	50.69	90.00	6.12
DeepPanoContext-3D	34.87	0.19	0.45	34.44	**47.59**	1.72	60.72	28.72	72.19	90.00	17.70
GroupFree	61.26	15.70	16.37	**65.92**	42.31	13.93	71.95	**31.02**	92.43	**68.59**	42.48
Ours	**67.94**	**16.20**	**17.18**	64.63	45.42	**18.47**	**72.12**	28.65	**94.62**	64.05	**44.53**

Method	wall mounter tv	loud speaker	console	fence	chest	standing tv	table lamp	speaker system	bathtub	plant	treadmill
Total3D-Pano	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.77	3.10	0.00
IM3D-Pano	0.08	0.00	0.00	0.00	0.00	0.00	3.17	0.00	10.26	12.69	0.00
DeepPanoContext	0.14	0.00	0.00	0.00	0.00	0.00	2.79	0.00	41.02	16.46	0.00
DeepPanoContext-3D	0.07	0.00	0.00	0.00	0.00	0.00	**17.79**	0.00	37.82	**50.14**	0.00
GroupFree	46.27	0.00	0.00	0.00	0.00	0.00	7.99	0.00	**100.00**	33.12	0.00
Ours	**69.30**	0.00	0.00	0.00	0.00	0.00	5.25	0.00	**100.00**	37.17	0.00

Method	washer	stool	trash can	stove	bed	office chair	towel rack	piano	shower	mAP	
Total3D-Pano	32.21	29.09	25.84	44.44	73.22	0.00	**50.00**	75.00	72.73	25.79	
IM3D-Pano	36.50	29.09	39.13	44.44	73.22	0.00	0.00	43.44	80.17	27.25	
DeepPanoContext	36.50	29.09	**66.23**	44.44	71.57	0.00	0.00	**100.00**	**100.00**	33.59	
DeepPanoContext-3D	**37.60**	29.09	54.32	**79.59**	72.40	0.00	0.00	**100.00**	**100.00**	39.26	
GroupFree	2.64	56.66	1.45	4.35	97.36	0.00	0.40	46.45	**100.00**	40.51	
Ours	12.80	**85.83**	4.38	6.54	**98.45**	0.00	0.00	34.79	**100.00**	**43.55**	

Table 7. 3D object detection performance per category on iGibson-Synthetic dataset, evaluated with mAP@0.15 IoU. The reason for categories with 0 mAP is that they exist in training scenes but do not exist in testing scenes, or vice versa.

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1

[4] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters*, 6 (2):1519–1526, 2021. 1

[5] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 3

[6] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 1

[7] Giovanni Pintore, Eva Almansa, Marco Agus, and Enrico Gobbetti. Deep3dlayout: 3d reconstruction of an indoor layout from a spherical panoramic image. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 3

[8] Danila Rukhovich and et al. Tr3d. In *ICIP*, 2023. 3

[9] Christoph Sager, Patrick Zschech, and Niklas Kuhl. labelCloud: A lightweight labeling tool for domain-agnostic 3d object detection in point clouds. *Computer-Aided Design and Applications*, 19(6):1191–1206, 2022. 1, 2

[10] Zhijie Shen and et al. Dopnet. In *CVPR*, 2023. 3

[11] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1

[12] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019. 3

[13] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021. 3

[14] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Led2-net: Monocular 360deg layout estimation via differentiable depth rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12956–12965, 2021. 3

[15] Shang-Ta Yang, Chi-Han Peng, Peter Wonka, and Hung-Kuo Chu. Panoannotator: A semi-automatic tool for indoor panorama layout annotation. In *SIGGRAPH Asia 2018 posters*, pages 1–2. 2018. 1, 2

[16] Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12632–12641, 2021. 1, 3

Method	cabinet	door	chair	curtain	lamp	rug	sofa	table	trash	tv
DeepPanoContext	35.33	6.78	47.04	13.6	12.15	4.49	26.87	73.34	39.59	4.86
DeepPanoContext-3D	52.49	11.42	**70.39**	32.38	20.02	9.10	30.13	**82.24**	**63.22**	12.19
Group-Free	59.56	42.21	52.83	**34.07**	19.65	32.90	80.59	51.47	44.64	**52.76**
Ours	**63.69**	**46.74**	54.02	30.41	**20.04**	**48.53**	**80.96**	46.42	51.53	47.82

Method	box	picture	basket	sink	fridge	window	pillow	towel	nightstand	bed
DeepPanoContext	16.41	0.01	0.00	80.78	80.50	0.02	2.48	0.00	4.00	7.58
DeepPanoContext-3D	32.71	4.83	0.48	**95.55**	92.57	15.22	4.10	0.00	4.13	15.79
GroupFree3D	16.25	4.01	0.36	47.53	**97.88**	24.96	23.67	0.00	26.39	89.33
Ours	**27.41**	**6.10**	**0.51**	55.51	95.69	**26.87**	**34.15**	0.00	**31.05**	**96.76**

Method	desk	toilet	mirror	stool	wall clock	mAP				
DeepPanoContext	0.00	4.69	0.00	0.00	0.87	18.45				
DeepPanoContext-3D	0.00	8.08	0.00	0.00	1.81	26.35				
GroupFree3D	**1.08**	56.11	0.61	0.24	0.27	34.38				
Ours	0.44	**59.30**	**3.70**	**0.29**	**11.83**	**37.59**				

Table 8. 3D object detection performance per category on ReplicaPano dataset, evaluated with mAP@0.15 IoU.

Figure 3. Qualitative comparisons on 3D object detection and scene reconstruction on iGibson-Synthetic. In the left two columns, we compare our object detection results with DeepPanoContext (DPC), DeepPanoContext with point cloud (DPC-3D), and ground truth in the panoramic view and bird's eye view. The color of the bounding boxes represents their categories. The third column shows the results of scene reconstruction, with two magnified object reconstruction results presented on the right-hand side.

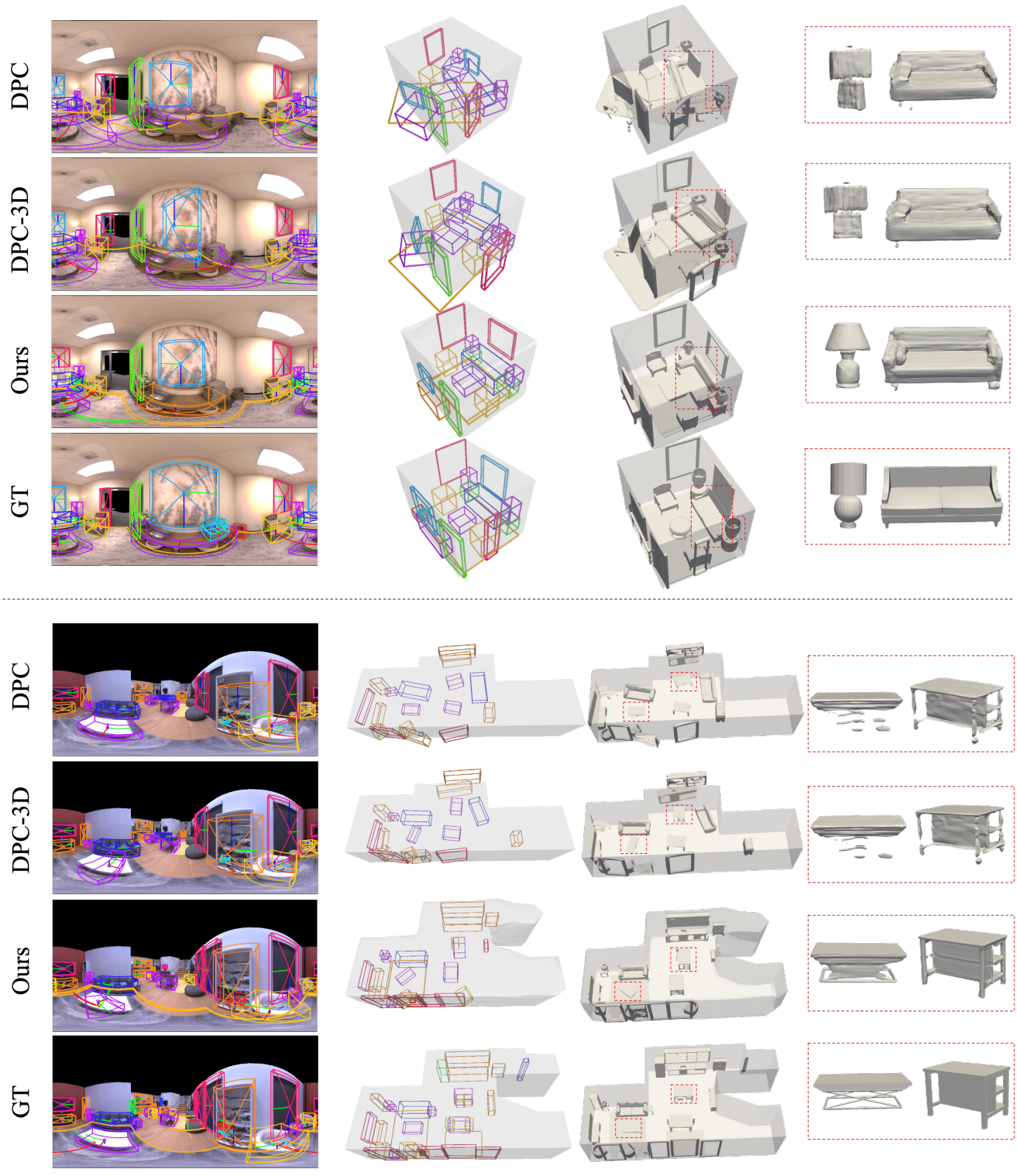

Figure 4. Qualitative comparisons on 3D object detection and scene reconstruction on ReplicaPano. The description of the figure is consistent with that of Fig. 3