DeepPanoContext: Panoramic 3D Scene Understanding with Holistic Scene Context Graph and Relation-based Optimization Supplementary Material

Cheng Zhang[1] Zhaopeng Cui[2*] Cai Chen[1] Shuaicheng Liu[1*] Bing Zeng[1] Hujun Bao[2] Yinda Zhang[3*]
[1] University of Electronic Science and Technology of China
[2] State Key Lab of CAD & CG, Zhejiang University [3] Google

In this supplementary material, we provide synthetic dataset examples, network architecture details, and implementation details. We also provide visualization of relation optimization, 3D detection performance on all categories, more qualitative results, more comparison on Structured3D, and discussion of failure cases.

A. Dataset Examples

Our synthetic dataset provides various ground truth along with the RGB panorama images, including 2D object bounding boxes/BFoVs, watertight scene/object meshes, oriented 3D object bounding boxes, and 3D room layout. Our synthetic panorama scene understanding dataset also provides depth maps and semantic/instance segmentation images, which can be used by others. Some examples of our panorama dataset are shown in Fig. C. We also show object crops collected from the panorama images and extra object-centric images rendered from object models used for single image object reconstruction in Fig. D. The data generation code is built upon iGibson [4] and fully customized for panorama images.

B. Implementation Details

Dealing with Panorama Image As mentioned in the main paper, to deal with the continuity of panorama images, we parameterize the 2D bounding box with Bounding FoV (BFoV) [2, 5], and extend the panorama boundary before running 2D detector. Moreover, we change the object orientation θ to be the yaw angle of the object in the cropped perspective image coordinate. Compared to directly estimating the orientation in the world frame as in Im3D [6] and Total3D [3], our representation is more intuitive because it explicitly codes the transformation from the camera coordinates to the world coordinates. When calculating bounding box projection term e^{bp} in relation optimization, we rotate the camera to each detected bounding box center then do the projection of 3D bounding boxes, which avoids cross-border situations.

RGCN Output		Loss Weight	
Symbol	Description	Symbol	Value
rr	Object-object/wall relative rotation relation	λ_{rr}	10
oa	Object-object/wall attachment relation	λ_{oa}	10
fa	Object-floor attachment relation	λ_{fa}	10
ca	Object-ceiling attachment relation	λ_{ca}	10
rd	Object-object relative distance relation	λ_{rd}	10
δ	3D bounding box center offset	λ'_{δ}	10
d	3D bounding box distance	λ'_{d}	10
s	3D bounding box size	λ'_{s}	10
θ	Object orientation	λ'_{θ}	10

Table A: RGCN outputs and loss weights of $\mathcal{L}_{RGCN}$ and $\mathcal{L}$.

Term		Weight	
Symbol	Description	Symbol	Value
oc	Object-object collision	λ^{oc}	1
wc	Object-wall collision	λ^{wc}	1
fc	Object-floor collision	λ^{fc}	1
cc	Object-ceiling collision	λ^{cc}	1
rr	Object-object/wall relative rotation relation	λ^{rr}	0.1
oa	Object-object/wall attachment relation	λ^{oa}	1
fa	Object-floor attachment relation	λ^{fa}	1
ca	Object-ceiling attachment relation	λ^{ca}	1
rd	Object-object relative distance relation	λ^{rd}	0.01
δ	3D bounding box center offset	λ^{rd}	0.0001
d	3D bounding box distance	λ^{d}	0.01
s	3D bounding box size	λ^{s}	1
θ	Object orientation	λ^{θ}	0.001
bp	3D bounding box projection	λ^{bp}	10

Table B: Terms in relation optimization and the weight of each term.

RGCN relation branch We design a relation branch for our RGCN to facilitate the relation estimation from the 512-dim representation vectors of object/relation nodes. We design the relation branch of RGCN as two-layer MLPs for each relation, which consist of a 256-dim FC layer, followed by a ReLU and Dropout layer with a drop factor of 0.5, and an output layer. The output layer is 1-dim for binary relations (*i.e.*, object-object/wall/floor/ceiling contact, inside or outside room, closer and farther to camera center between a pair of objects), and 8-dim for multi-class relations (*i.e.*, the

*Corresponding author

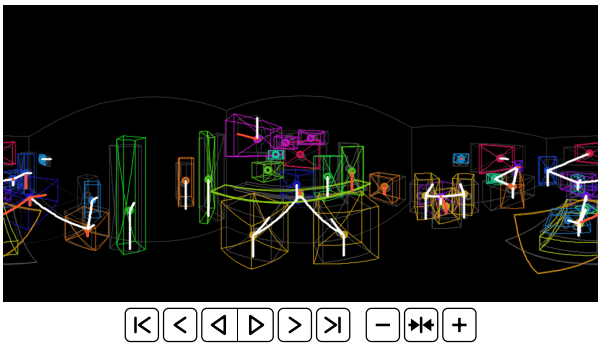

Figure A: Visualization of our proposed relation optimization. A PDF reader like **Adobe Acrobat Reader / KDE Okular** might be needed for displaying animated sequences. We also include more animation as a video along with this pdf file. The ground truth object bounding boxes are visualized with gray color for reference, while the current states are colorized. The attachment relations among objects, walls, floor, and ceiling are indicated by thick white lines, while the collisions are in red.

angular difference between object and object/wall).

Hyper parameters For the weights of $\mathcal{L}_{ODN}$, we refer to Total3D [3] for detailed settings. For the loss weights of $\mathcal{L}_{RGCN}$ and the joint loss $\mathcal{L}$, we show the description of each output and its corresponding loss weight in Tab. A. For relation optimization, we weight each term by its confidence and importance. For example, 2D observations should be more confident, and collision terms should be weighted more if we consider physically plausible object poses important. We show the description of each term and its corresponding weight in Tab. B. In optimization, we use a gradient descend optimizer and set the learning rate to 1, steps to 100, and momentum to 0.9.

Training All the borrowed networks (*i.e.*, Mask-RCNN, HorizonNet, ODN, LIEN, LDIF) are fine-tuned individually on our proposed dataset. Specifically, Mask-RCNN is fine-tuned from the weights pre-trained on COCO dataset, with batch size of 8 and learning rate of 2e-3 for 1e5 steps. HorizonNet is fine-tuned from the weights pre-trained on Structured3D dataset, with batch size of 6 and learning rate of 1e-4 for 50 epochs. ODN is fine-tuned from the weights pre-trained on SUN RGB-D, with batch size of 6 and learning rate of 1e-4 for 15 epochs. LIEN and LDIF are fine-tuned from the weights pre-trained on Pix3D, with batch size of 24 and learning rate of 2e-4 for 100 epochs. To make a fair comparison, all variation of Total3D [3] and Im3D [6] including the perspective and panorama version are also fine-tuned on our proposed dataset following the above process. For Total3D-Pers and Im3D-Pers, the ODN and Scene Graph Convolutional Network (SGCN) are fine-tuned and tested with detection results obtained from split views. In addition, MGNet used by Total3D is fine-tuned from the

Metric	Total3D		Im3D	
	w/o. RO	w. RO	w/o. RO	w. RO
mAP (57 categories, ↑)	25.79	32.46	27.25	33.54
avg col (↓)	3.41	0.89	2.62	0.90

Table C: The improvement of RO on different methods. The improvement of 3D object detection is evaluated with mAP of all 57 categories and physical violation is evaluated with average collision times per scene.

Methods (Pano)	Initial Estimation	Object Reconstruction	GCN	RO	Total
Total3D	0.66 (Mask R-CNN, HorizonNet, ODN)	0.23 (MGN)	-	-	0.89
Im3D		5.92 (LIEN+LDIF)	0.03 (Scene GCN)	-	6.62
Ours			0.06 (Relation-based GCN)	4.74	11.38

Table D: Efficiency comparison. We use average time per scene in seconds to compare efficiency of different methods and modules (tested on a single GTX 1080Ti).

weights pre-trained on Pix3D, with batch size of 16 and learning rate of 1e-4 for 100 epochs. To train our proposed RGCN, we generate the attachment relation ground truth by doing collision detection with a tolerance of 0.1m (*i.e.*, before collision detection, we expand the bounding box by 0.05m) on the ground-truth 3D object bounding boxes and the estimated layout walls. The other relations are calculated according to their definition directly. We first train our RGCN with only the pose refinement branch, with batch size of 16 and learning rate of 1e-4 for 35 epochs. Then we fine-tune it with relation estimation branch for 20 epochs using the same settings. Finally, we do an end-to-end training of ODN and RGCN with RO, with batch size of 1 and learning rate of 1e-5 for 10 epochs.

C. Visualization of Relation Optimization

To visualize the process of our proposed relation optimization, we present an animation in Fig. A. For demonstration, we add random noises to the ground truth object poses as the initial state, which simulates the inaccuracy of the initial pose estimation. We then use the relation generated from the ground truth poses to optimize the current poses (colorized) using our proposed method. We observe that as the optimization goes on, the position and orientation of the objects become closer to the ground truth, while the collisions are gradually resolved.

D. 3D Detection mAP on all 57 categories

In Tab. 1 of the main paper, we show 3D object detection results for 11 common categories. Here we show a complete quantitative evaluation on all 57 categories in Tab. G. Same as the conclusion made in the main paper, our method outperforms the SoTA with a large margin.

Method (Pano)	door	picture	table	sofa	chair	window	bed	bottom cabinet	chest
Total3D	28.65	0.06	38.83	31.64	23.71	4.78	74.09	37.08	62.07
Im3D	37.59	0.14	**49.47**	**37.24**	29.34	6.35	77.66	45.18	70.03
Ours (w/o. RO)	54.74	0.69	48.39	36.05	29.85	**13.49**	81.13	48.33	72.08
Ours (Full)	**57.73**	**1.24**	49.10	37.02	**29.95**	12.28	**81.15**	**48.76**	**74.26**

Method (Pano)	sink	fridge	bathtub	shelf	mirror	toilet	counter	standing tv	mean
Total3D	28.24	68.82	69.36	10.36	0.04	19.88	19.17	2.12	30.52
Im3D	**28.57**	71.39	**73.93**	9.78	0.92	15.04	19.17	2.52	33.78
Ours (w/o. RO)	27.43	**73.35**	**73.93**	**15.84**	1.47	32.87	19.17	9.61	37.55
Ours (Full)	27.93	**73.35**	**73.93**	15.76	**3.19**	**65.54**	19.17	**13.20**	**40.21**

Table E: 3D object detection comparison on Structured3D. We evaluate on the 17 iGibson categories mapped from 20 Structured3D categories and use mean average precision (mAP) with the threshold of 3D bounding box IoU set at 0.15 as the evaluation metric.

(a) Input (b) 3D Detection (c) Reconstruction

Figure B: Qualitative results of our model on Structured3D.

E. More Qualitative Comparison on 3D Detection and Scene Reconstruction

In Sec. 4.1 of the paper, we show qualitative comparisons on 3D detection and reconstruction. Here we provide more results in Fig. E. Compared to the SoTA methods [3, 6], our method produces significantly better 3D detection and reconstruction results. From the 3D detection and reconstruction results in panorama view, we observe that our method generates more accurate projections of reconstructed objects (*e.g.*, the mirror of (a), the sofa of (b) and (d), the door of (c)). From the 3D detection results in Bird's Eye View, we can see that our method generates more reasonable and physically plausible object poses (*e.g.*, (c), (e) have less object-wall collision and better rotation relations with walls).

F. Would RO Improve Other Methods?

In order to further evaluate the proposed relation optimization, we apply our RO on Total3D and Im3D using our predicted relation and their final results, and show the results in Tab. C. We can see that both methods still significantly benefit from the RO, which demonstrates that our RO is effective and robust to different initial estimates.

Parameter	Value	Parameter	Value
λ^{rd}	0.0040	λ^{oc}	0.0157
λ^{d}	0.1404	λ^{wc}	0.2625
λ^{s}	6.0502	λ^{fc}	0.3182
λ^{θ}	0.0003	λ^{cc}	0.2036
λ^{bp}	0.2895	learning rate	0.0124

Table F: Auto-searched hyperparameters used on Structured3D, including weights of relation optimization terms and learning rate of relation optimization.

G. Run-time Efficiency

The efficiency comparison is shown in Tab. D. It is worth mentioning that implicit representation LDIF and RO are all implemented with PyTorch, and can be further optimized, *e.g.*, using CUDA, to improve the efficiency.

H. Experiment on Structured3D

Since Structured3D provides the ground truth of object pose and layout, we can train our model up to RGCN. Due to the lack of mesh ground truth, we load the object reconstruction model with weights trained on iGibson. Furthermore, since the object reconstruction model requires category label as input, we map the object categories from Structured3D to iGibson. We found overlapping categories between two datasets, which ends up with 20 structure3D categories mapped to 17 iGibson categories. Specifically, "cabinet", "bookshelf", "desk", "shelves", "dresser", "floor mat", "television", "box", "nightstand" in Structure3D are mapped to "bottom cabinet", "shelf", "table", "shelf", "bottom cabinet", "carpet", "standing tv", "chest", "chest" in iGibson, and others are mapped with the same category name. It is also worth mentioning that the bounding box GT of objects in Structured3D is not accurate or physical plausible, which makes it difficult to produce rich relation GT and to better refine the object poses with observation and collision terms. So the weights of relation optimization terms

need to be tuned to match the condition. Specifically, we fix the weights of relation terms and auto-search the learning rate of gradient descend optimizer and other weights of relation optimization terms around the original settings used on iGibson. In summary, we train object detection on overlapping categories and set weights of RO terms with auto-search [1]. The auto-searched weights are shown in Tab. F. Qualitative results are shown in Fig. B. We can see that our method performs well with good layout, pose and shape estimation although there is no ground truth for shapes. We compare 3D object detection against existing methods quantitatively in Tab. E. The results show that our method still outperforms SoTA methods significantly, and RO plays a big role in improving the mAP.

I. Failure Cases

We show failure cases in Fig. F. One scenario that our pipeline fails is when heavy occlusion happens (*i.e.*, one of the doors on the right in (d), the second door on the left in (a)), which tends to shrink the size of the object in order to favor the projection term with the partial 2D observation. A possible solution might be to understand the occlusion and learn the mask behind occluder. Another scenario is when the 2D detector has multiple detection results on a single object (*i.e.*, the wardrobe on the right in (a), the sofa on the right in (b), the drawer on the left in (c)), which lead to two overlapped object reconstructions in the same place but not sufficient to trigger non-maximum suppression. This might be solved by refining the category prediction of the 2D detector in the RGCN, which will presumably fix detected object categories with mistakes (or set reduplicated object to void) with a better understanding of the 3D scene context. The last scenario is when HorizonNet fails to generate layouts for rooms that don't satisfy the Manhattan-world assumption (*i.e.*, the wall on the left side in (e)), our pipeline will fail to optimize the object pose based on the wrong wall orientation. Also when object-wall rotation relation is estimated badly (*i.e.*, the window in (b)), the orientation cannot be optimized properly.

References

[1] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com. 4

[2] Shih-Han Chou, Cheng Sun, Wen-Yen Chang, Wan-Ting Hsu, Min Sun, and Jianlong Fu. 360-indoor: Towards learning real-world objects in 360deg indoor equirectangular images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 845–853, 2020. 1

[3] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2, 3

[4] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Shyamal Buch, Claudia D'Arpino, Sanjana Srivastava, Lyne P Tchapmi, et al. igibson, a simulation environment for interactive tasks in large realisticscenes. *arXiv preprint arXiv:2012.02924*, 2020. 1

[5] Wenyan Yang, Yanlin Qian, Joni-Kristian Kämäräinen, Francesco Cricri, and Lixin Fan. Object detection in equirectangular panorama. In *Int. Conf. Pattern Recog.*, pages 2190–2195. IEEE, 2018. 1

[6] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8833–8842, June 2021. 1, 2, 3

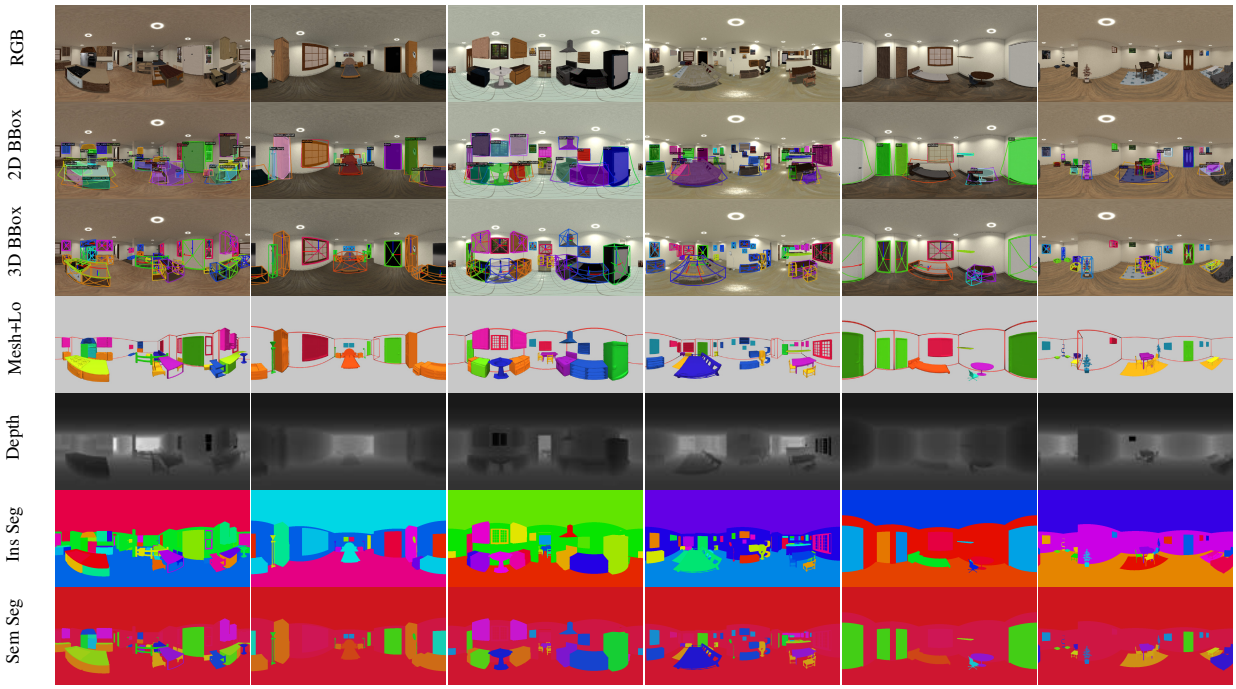

Figure C: Samples of our proposed panorama 3D scene understanding dataset.

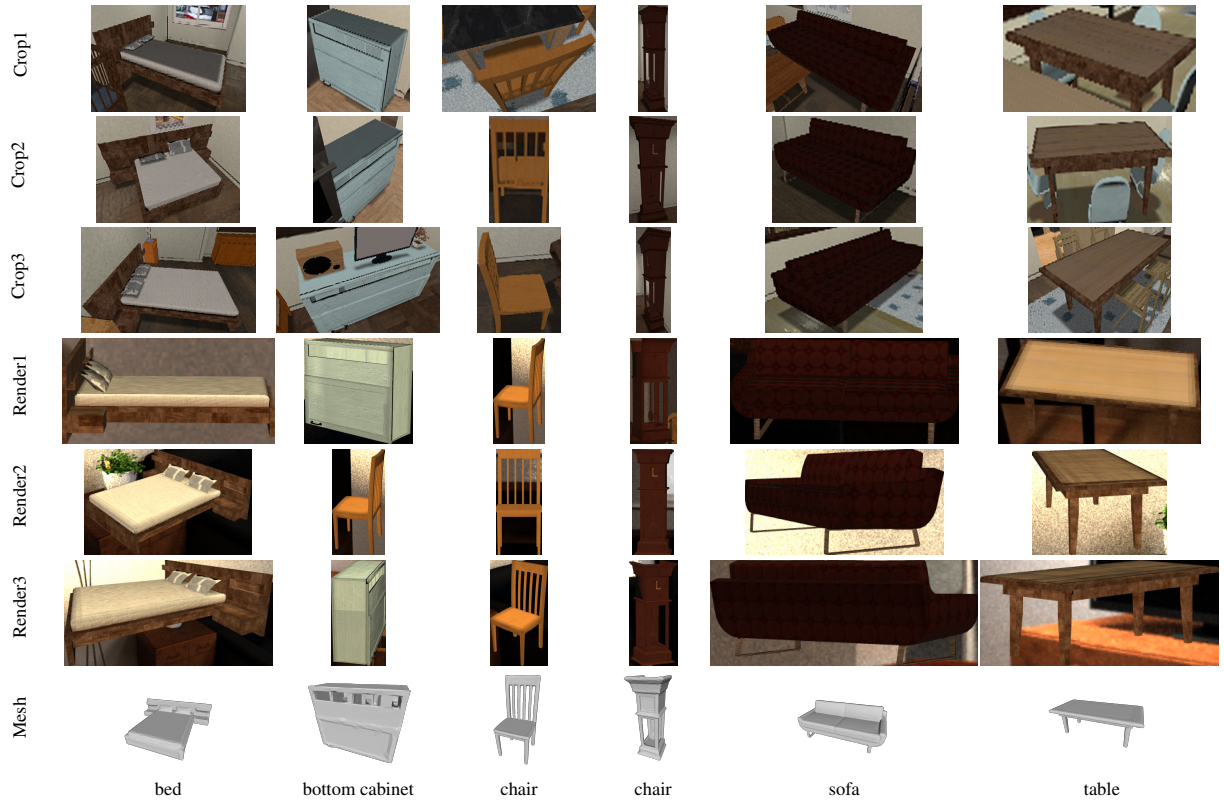

Figure D: Samples of dataset used for single image object reconstruction.

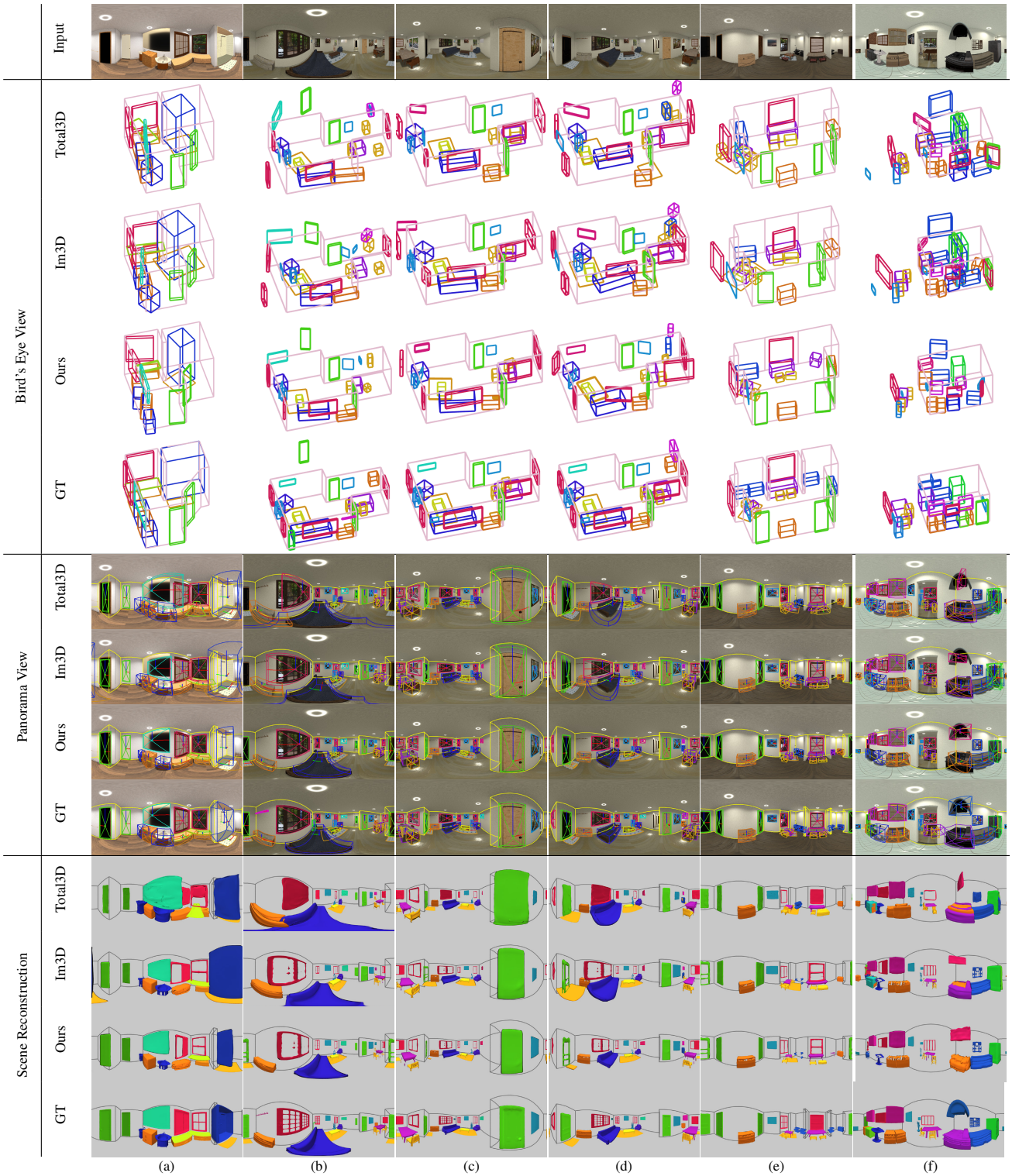

Figure E: More Qualitative comparison on 3D object detection and scene reconstruction.

Method	chair	sofa	table	fridge	sink	door	floor lamp	bottom cabinet	top cabinet	sofa chair	dryer
Total3D-Pers	13.71	68.06	30.55	36.02	69.84	11.88	12.57	35.56	19.19	64.29	41.36
Total3D-Pano	20.84	69.65	31.79	43.13	68.42	10.27	16.42	34.42	20.83	62.38	33.78
Im3D-Pers	30.23	**75.23**	44.16	52.56	76.46	14.91	9.99	45.51	23.37	**80.11**	53.28
Im3D-Pano	33.08	72.15	37.43	70.45	75.20	11.58	6.06	43.28	18.99	78.46	41.02
Ours (w/o. RO)	**33.57**	75.18	38.65	71.97	**80.66**	19.94	18.29	50.67	29.05	79.42	**60.07**
Ours (Full)	27.78	73.96	**46.85**	**74.22**	75.29	**21.43**	**20.69**	**52.03**	**50.39**	77.09	59.91

Method	window	carpet	picture	oven	bottom cabinet no top	counter	dish washer	shelf	coffee table	mirror	toilet
Total3D-Pers	2.92	0.05	0.01	31.33	34.40	0.78	43.54	10.93	39.72	0.11	90.00
Total3D-Pano	3.07	0.05	0.02	29.81	32.48	1.11	48.39	9.57	49.52	0.64	90.00
Im3D-Pers	3.52	0.12	0.00	31.28	**47.45**	**2.60**	51.47	15.01	59.02	0.81	90.00
Im3D-Pano	3.42	0.01	0.01	29.06	44.79	1.34	43.80	15.41	56.82	0.16	90.00
Ours (w/o. RO)	6.94	0.12	0.03	32.52	46.42	1.83	59.78	15.58	**61.17**	2.42	90.00
Ours (Full)	**9.56**	**0.65**	**0.21**	**34.50**	44.17	1.25	**63.19**	**22.65**	50.69	**6.12**	90.00

Method	wall mounted tv	loud speaker	console table	fence	chest	standing tv	table lamp	speaker system	bathtub	plant	treadmill
Total3D-Pers	**0.52**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.77	11.48	0.00
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-Pers	0.03	0.00	0.00	0.00	0.00	0.00	6.06	0.00	10.26	10.34	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
Ours (w/o. RO)	0.24	0.00	0.00	0.00	0.00	0.00	**10.53**	0.00	10.26	8.35	0.00
Ours (Full)	0.14	0.00	0.00	0.00	0.00	0.00	2.79	0.00	**41.02**	**16.46**	0.00

Method	washer	stool	trash can	stove	bed	office chair	shower	towel rack	piano	mAP	
Total3D-Pers	35.06	29.09	24.45	44.44	71.87	0.00	**100.00**	25.00	55.56	25.11	
Total3D-Pano	32.21	29.09	25.84	44.44	**73.22**	0.00	72.73	**50.00**	75.00	25.79	
Im3D-Pers	**36.50**	29.09	22.02	44.44	**73.22**	0.00	81.82	**50.00**	83.33	29.86	
Im3D-Pano	**36.50**	29.09	39.13	44.44	**73.22**	0.00	80.17	0.00	43.33	27.25	
Ours (w/o. RO)	**36.50**	29.09	31.15	44.44	71.57	0.00	81.82	0.00	**100.00**	30.91	
Ours (Full)	**36.50**	29.09	**66.23**	44.44	71.57	0.00	**100.00**	0.00	**100.00**	**33.59**	

Table G: 3D object detection comparison on full 57 categories. Some categories existing in training scenes do not exist in testing scenes, or vice versa, which is the main reason for some of the 0 mAP cases.

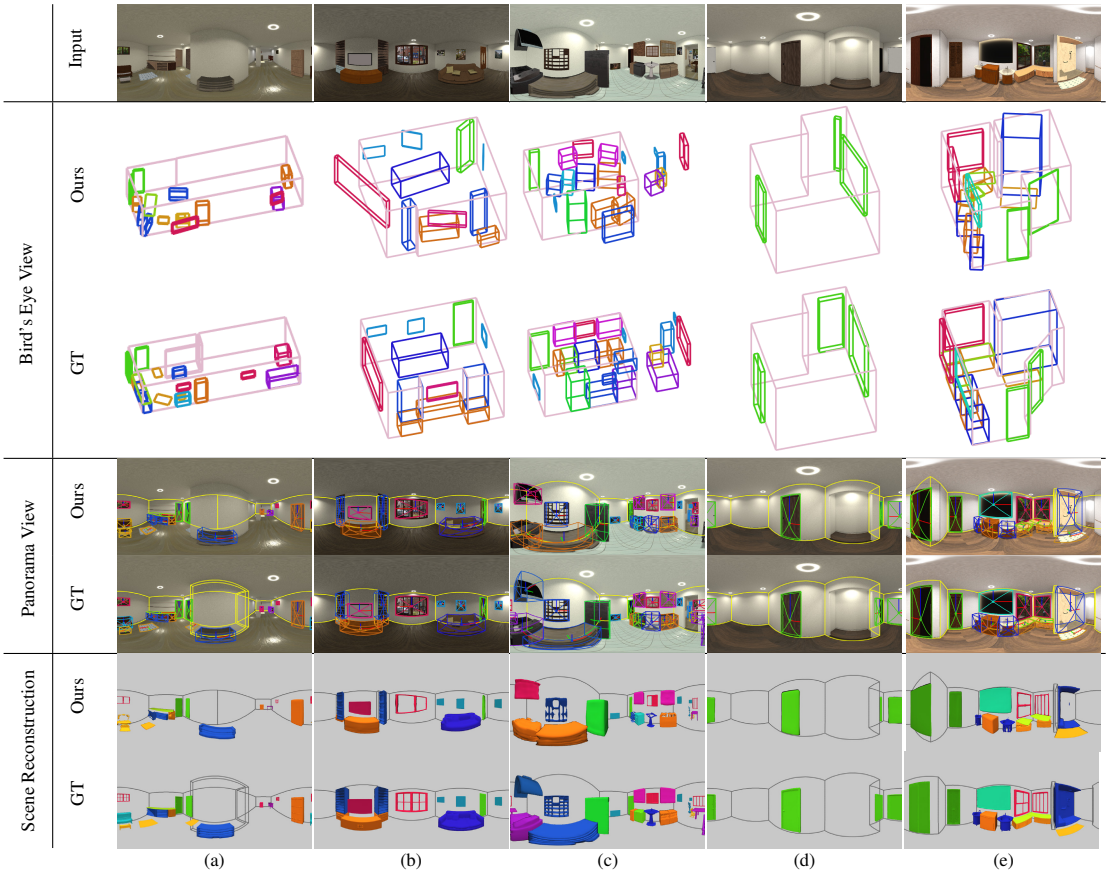

Figure F: Failure cases.