iBARLE: imBalance-Aware Room Layout Estimation - Supplementary

Taotao Jing[1], Lichen Wang[2], Naji Khosravan[2], Zhiqiang Wan[2], Zachary Bessinger[2], Zhengming Ding[1], Sing Bing Kang[2]
[1]Tulane University [2]Zillow Group
{tjing,zding1}@tulane.edu
{lichenw,najik,zhiqiangw,zacharybe,singbingk}@zillowgroup.com

We first describe important attributes of ZInD [1] and how we use ZInD in our work. We then show more ablation analysis results, to give insight on the contribution of each module in our proposed model. Finally, more comparative qualitative experimental results are shown to support the claim of our system being state-of-the-art.

1. More Experimental Results

We performed additional experiments on the Structured3D (S3D) dataset [4] to showcase the effectiveness of iBARLE in handling complex indoor scenes with furniture. Results in Table R1 demonstrate significant improvements in both "overall" and "group-wise average" 2D IoU metrics. We also conducted cross-domain zero-shot evaluations (train on the ZinD [6] and test on S3D), showing promising results in handling domain-shift of structural/visual variations.

Table R1. Statistics of S3D and Layout Estimation Results

Corner Number			4	6	8	10+	Avg.	Overall
Number of Training Samples			11,507	3,136	1,287	2,231	–	18,161
Number of Test Samples			1,063	289	130	202	–	1,684
Trained on S3D [4]								
2D IoU ↑	LGT-Net [19]	*CVPR*' 22	95.06	91.59	90.66	84.44	90.44	92.85
	iBARLE	Ours	**95.77**	**93.99**	**93.74**	**91.01**	**93.63**	**94.74**
Trained on ZinD [6] (Cross-domain Zero-shot)								
2D IoU ↑	LGT-Net [19]	*CVPR*' 22	85.67	78.10	72.67	50.24	71.67	79.12
	iBARLE	Ours	**86.47**	**78.73**	**73.54**	**52.46**	**72.80**	**80.06**

Moreover, we compared our model with baselines using a sampling strategy to balance the S3D training data. The results in Table R2 showed that the sampling strategy did not effectively contribute due to the significant imbalance and led to overfitting to the over-sampled minority groups data.

Table R2. Layout Estimation on S3D Dataset (2D IoU ↑)

Corner Number		4	6	8	10+	Avg.	Overall
LGT-Net [19]	*CVPR*' 22	95.06	91.59	90.66	84.44	90.44	92.85
LGT-Net + Sampling	*CVPR*' 22	89.78	88.54	87.70	85.89	87.98	88.94
iBARLE	Ours	**95.77**	**93.99**	**93.74**	**91.01**	**93.63**	**94.74**

2. Zillow Indoor Dataset

Zillow Indoor Dataset (ZInD) [1] is the largest indoor panorama image dataset with layout annotations for real residential homes. Specifically, there are $71,474$ panoramas from $1,524$ real unfurnished homes. The annotations include 2D/3D layouts, 2D floor plans, camera pose, and openings such as windows and doors. While layouts featured in other indoor datasets are mostly simple cuboid or Manhattan layouts, ZInD has a real-world distribution of layout complexities.

ZInD can be split into different subgroups based on layout attributes. As shown in Figure 1, floor plans can be separated into groups with different number of corners and room types (i.e., Manhattan-L, Non-Manhataan, etc). However, since the dataset is extremely imbalanced across different subgroups, the layout estimation model trained on such training data will be more reliable for those groups with sufficient samples, and less reliable for samples from the minority subgroups. This is the primary motivation of our imbalance-aware room layout estimation work.

We focus on three kinds of data splits:

- *Number of corners.* The whole dataset can be split based on the number of corners, as shown in Fig. 4 in the main paper. 43% of the dataset are rooms with 4 corners while 21% are with 6 corners. However, rooms with 9 corners and 7 corners constitute only 2% and 4% of the dataset, respectively. Rooms with $10+$ corners occupy 14% of ZInD, and they are substantially more complex, making them very challenging for room layout estimation.

- *Room types.* ZInD provides room type labels that include "Cuboid", "Manhattan-L", "Manhattan-General", and "Non-Manhattan". Many prior layout estimation solutions handle mostly simple cuboid rooms [2]. In this work, we explore the different performances across various room types with imbalanced numbers of training data available.

Figure 1. Selected examples from ZInD dataset [1] with various layout attributes.

- *Room position: Primary versus Secondary.* ZInD is the first large-scale indoor dataset containing multiple panoramas of the same room captured at different locations. Each panorama location is labeled either "Primary" and "Secondary". The "Primary" label is based on the perception that it is easier to use for layout estimation, and is usually close to the center of the room. On the other hand, "Secondary" panoramas are typically near walls or corners, which makes the layout estimation more challenging. As shown in Figure 4 in the manuscript, although the "Primary" and "Secondary" groups seem balanced, 52% V.S. 48%, "Secondary" panoramas location is much more diverse compared to the "primary" group. This makes layout estimation using data from the "Secondary" group more difficult.

3. Ablation Analysis

To demonstrate the importance of each module in the proposed model, we show more ablation analysis results in Tables 3, 4, and 5. Specifically, "Basic" denotes the results produced by the basic model without the AVC, CSMix, and gradient-based corner and occlusion boundaries constraint modules. "w/ AVC only", "w/ CSMix only", and "w/ gradient only" are the results obtained by adding each module to the basic framework. "Ours (Complete)" denote the results achieved by our complete iBARLE model. The results show the necessity of each module in our proposed system.

4. Qualitative Analysis

In this section, we select more examples (from each subgroup) from ZInD and compare our layout estimation results with those by LGT-Net [2] and LED^2-Net [3]. Figures 3-8 show results of representative samples from groups with different numbers of corners. Figures 9 and 10 compare performance for the "Primary" and "Secondary" panoramas, respectively. Results for the room type group are shown in Figures 11-14 ("Cuboid", "Manhattan-L", "Manhattan-General", and "Non-Manhattan", respectively).

Factors that make room shape estimation more complex include: there are many corner numbers, the shape is non-Manhattan, and the camera location is at a challenging part of the room. These factors result in occlusions, with different parts of the room unseen by the panorama. This can be seen in Figures 9 and 12. Such occlusion-based challenges do not exist in simple shape layouts that are used in prior work. Our proposed gradient-based corners and occlusions boundaries constraint is designed to manage such cases.

5. Visualization of Mixup Samples

We show one *CSMix* augmented sample as below. The *CSMix* module may have limitations with highly irregular or structurally complex layouts, which could lead to unrealistic or invalid results. Balancing the benefits of mixup augmentation and its limitations is crucial, considering dataset and task variations.

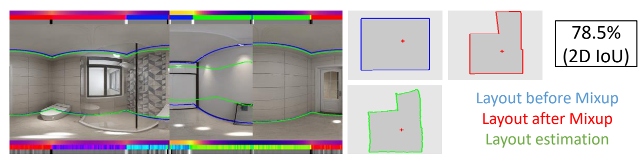


Figure 2. Selected Mix-up samples visualization.

References

[1] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow Indoor Dataset: Annotated floor plans with 360 panoramas and 3D room layouts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2133–2143, 2021. 1, 2

[2] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. LGT-Net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1654–1663, 2022. 1, 2

[3] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. LED2-Net: Monocular 360 layout estimation via differentiable depth rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12951–12960, 2021. 2

[4] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of the European Conference on Computer Vision*, pages 519–535. Springer, 2020. 1

Table 3. Ablation study of the contribution of each module in the proposed framework - split by number of corners

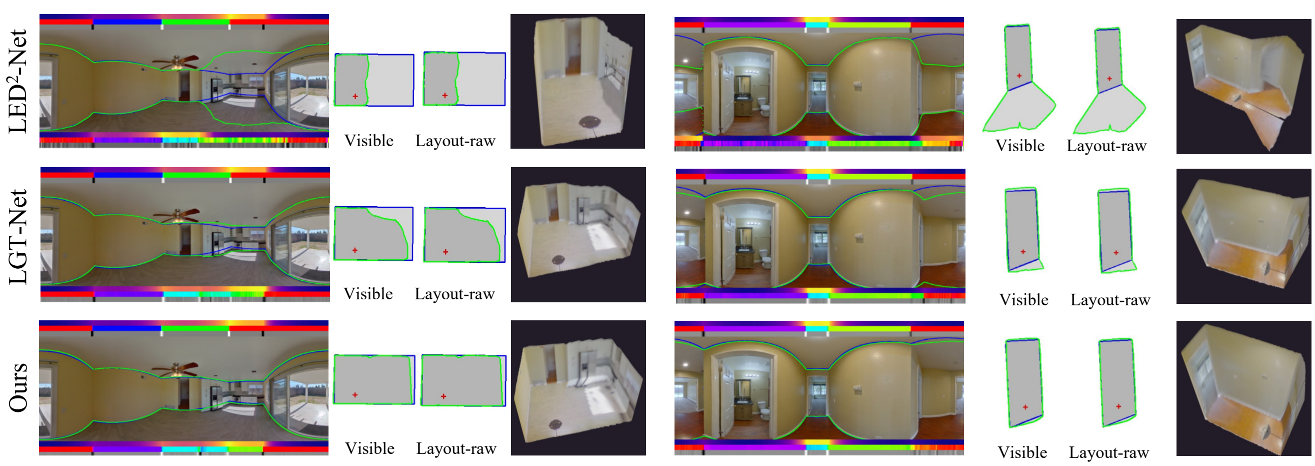
Corner	Basic				w/ AVC only				w/ CSMix only				w/ gradient only				Ours (Complete)			
Number	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1
4	87.21	85.37	**0.17**	**0.94**	87.79	86.00	0.18	**0.94**	**88.23**	**86.41**	0.18	**0.94**	88.07	86.21	0.18	**0.94**	88.22	86.38	0.18	**0.94**
5	85.76	83.44	0.22	0.93	87.48	85.31	**0.20**	**0.94**	87.50	85.26	**0.20**	0.93	87.44	85.35	**0.20**	0.93	**87.83**	**85.74**	**0.20**	0.93
6	83.50	81.66	0.21	0.93	85.16	83.28	0.20	0.94	84.95	83.08	0.20	0.94	85.43	83.54	**0.19**	**0.94**	**85.50**	**83.57**	**0.19**	**0.94**
7	79.68	77.22	0.28	0.91	79.97	76.78	0.26	0.91	**80.11**	**77.29**	**0.24**	0.91	79.78	77.03	0.25	0.91	79.62	76.92	0.25	**0.92**
8	80.13	77.98	0.23	0.92	80.60	78.51	**0.20**	0.93	80.67	78.45	**0.20**	0.93	**81.21**	**79.08**	**0.20**	**0.94**	80.69	78.55	**0.20**	**0.94**
9	80.39	78.17	0.26	0.92	80.58	78.44	**0.23**	**0.93**	80.88	78.67	**0.23**	**0.93**	80.93	78.53	**0.23**	**0.93**	**81.14**	**78.75**	**0.23**	**0.93**
10+	75.21	72.26	0.29	0.90	75.78	73.11	0.26	**0.92**	74.97	72.43	0.26	**0.92**	75.56	72.79	0.26	**0.92**	**76.16**	**73.39**	**0.25**	**0.92**
Avg.	81.70	79.44	0.24	0.92	82.48	80.21	0.22	**0.93**	82.47	80.23	0.22	**0.93**	82.63	80.36	0.22	**0.93**	**82.74**	**80.47**	**0.21**	**0.93**

Table 4. Ablation study of the contribution of each module in the proposed framework - split by camera pose

Camera	Basic				w/ AVC only				w/ CSMix only				w/ gradient only				Ours (Complete)			
Pose	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1
Primary	86.23	84.41	**0.19**	**0.94**	87.13	85.32	**0.19**	**0.94**	87.34	85.52	**0.19**	**0.94**	87.47	85.60	**0.19**	**0.94**	**87.72**	**85.85**	**0.19**	**0.94**
Secondary	81.57	79.33	0.22	0.93	82.53	80.35	**0.20**	**0.94**	82.48	80.30	**0.20**	0.93	**82.67**	**80.48**	**0.20**	0.93	82.63	80.44	**0.20**	0.93
Avg.	83.90	81.87	0.21	0.93	84.83	82.84	0.20	**0.94**	84.91	82.91	0.20	**0.94**	85.07	83.04	0.20	**0.94**	**85.18**	**83.15**	**0.19**	**0.94**

Table 5. Ablation study of the contribution of each module in the proposed framework - split by room type

Room	Basic				w/ AVC only				w/ CSMix only				w/ gradient only				Ours (Complete)			
Type	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1	2DIoU	3DIoU	RMSE	δ_1
Cuboid	87.54	85.69	**0.17**	**0.94**	88.16	86.36	0.18	**0.94**	88.58	86.74	0.18	**0.94**	88.47	86.60	0.18	**0.94**	**88.62**	**86.76**	0.18	0.94
Manhattan-L	83.29	81.43	0.21	0.93	84.86	82.99	0.20	**0.94**	84.63	82.77	0.20	**0.94**	**85.14**	**83.25**	**0.19**	**0.94**	85.13	83.21	**0.19**	**0.94**
Manhattan-General	78.19	75.90	0.25	0.92	78.60	76.63	**0.21**	**0.93**	78.41	76.42	**0.21**	**0.93**	**79.05**	**77.01**	**0.21**	**0.93**	78.90	76.89	**0.21**	**0.93**
Non-Manhattan	80.83	78.33	0.25	0.92	81.92	79.19	**0.23**	**0.93**	81.78	79.10	**0.23**	**0.93**	81.68	79.00	0.24	**0.93**	**82.08**	**79.36**	**0.23**	**0.93**
Avg.	82.46	80.34	0.22	0.93	83.39	81.29	0.21	**0.94**	83.35	81.26	0.21	0.93	83.58	81.46	0.20	0.93	**83.68**	**81.55**	**0.20**	**0.94**



(a) (b)

Figure 3. Case study: corner number = 4.

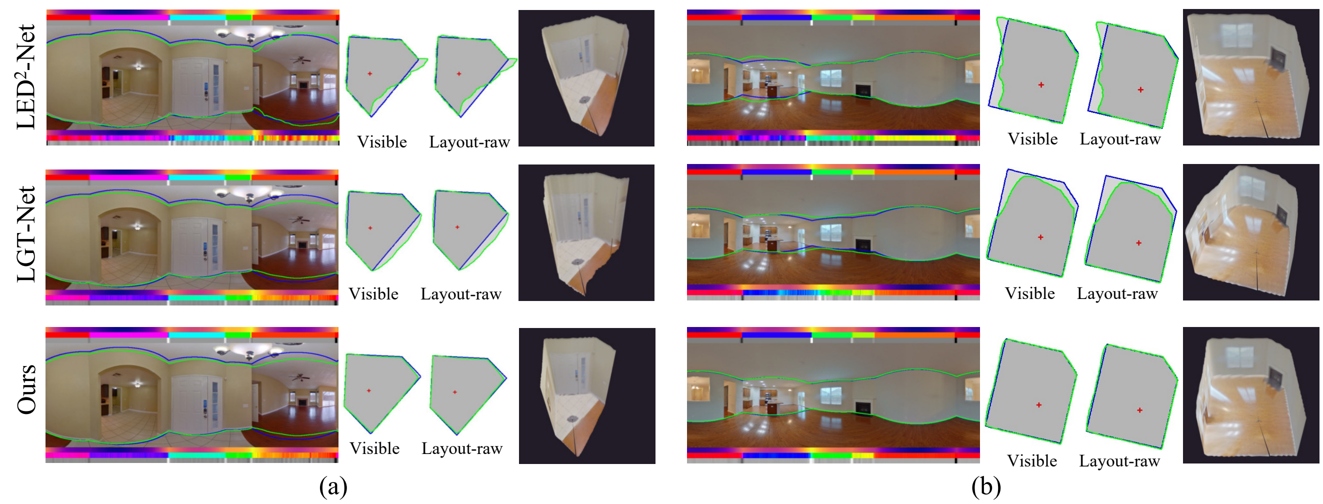


Figure 4. Case study: corner number = 5.

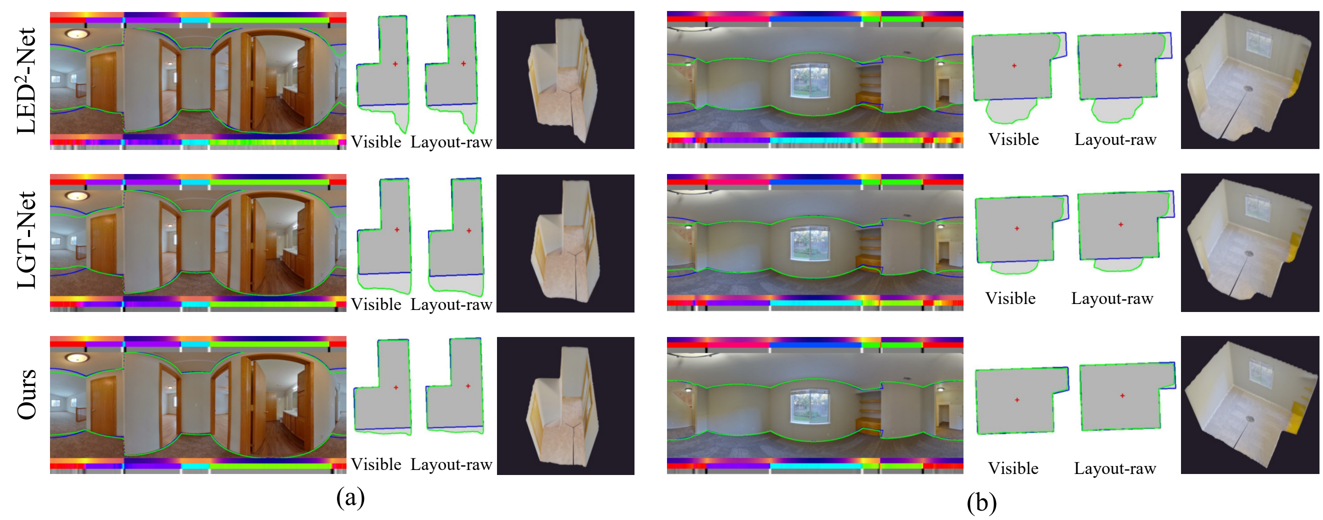


Figure 5. Case study: corner number = 6.

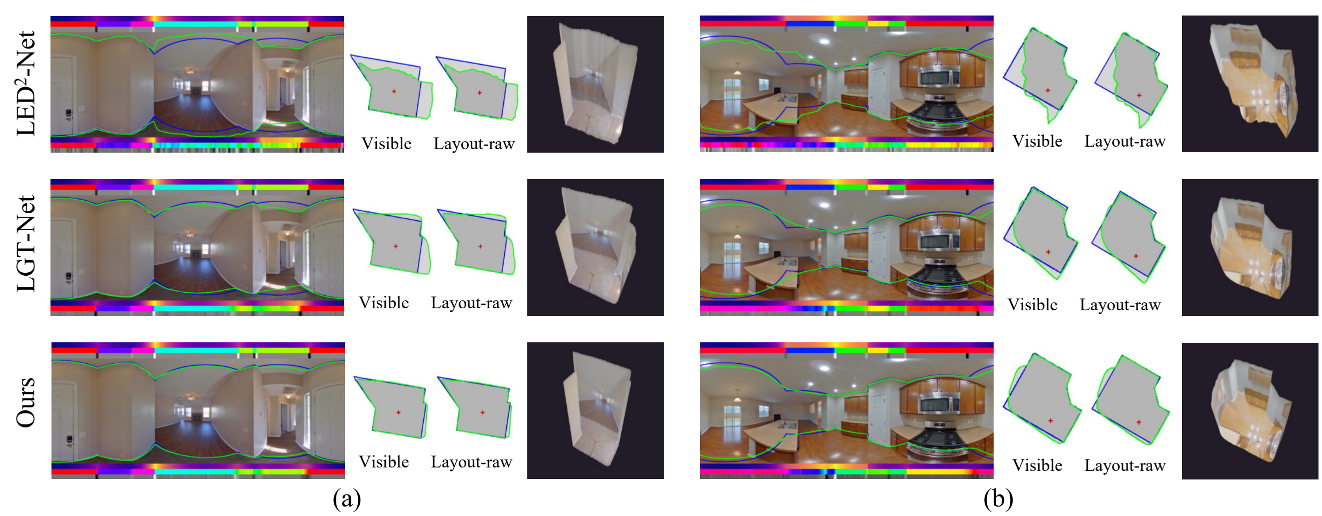


Figure 6. Case study: corner number = 7.

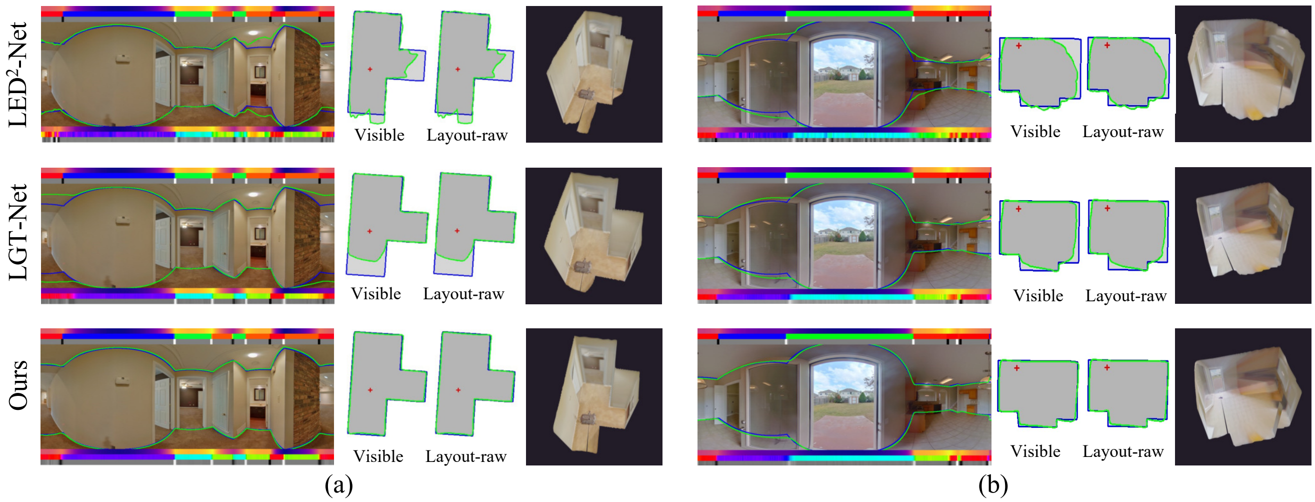


Figure 7. Case study: corner number = 8.

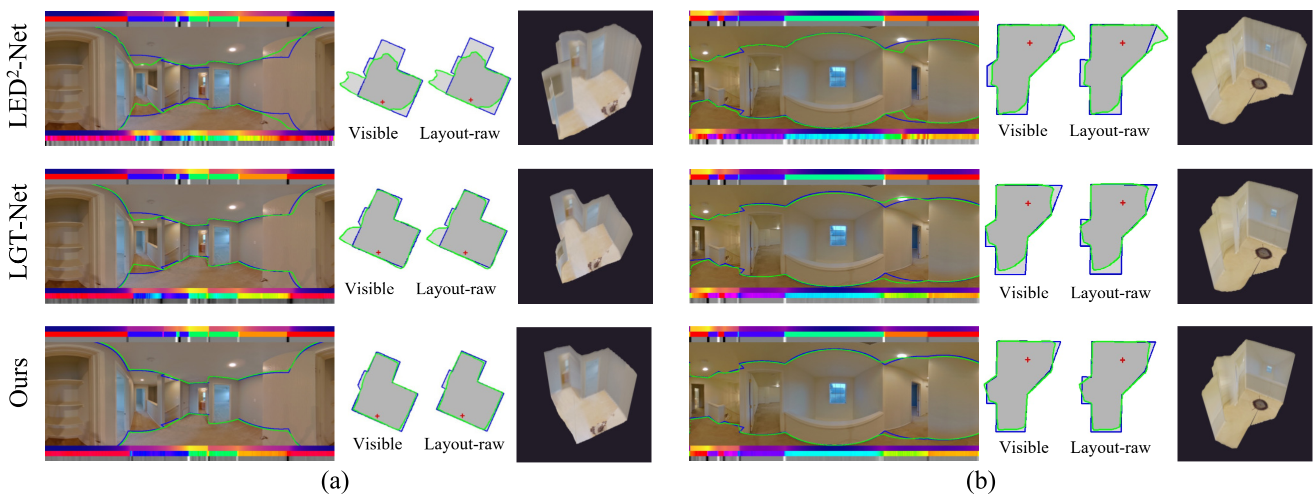


Figure 8. Case study: corner number = 10.

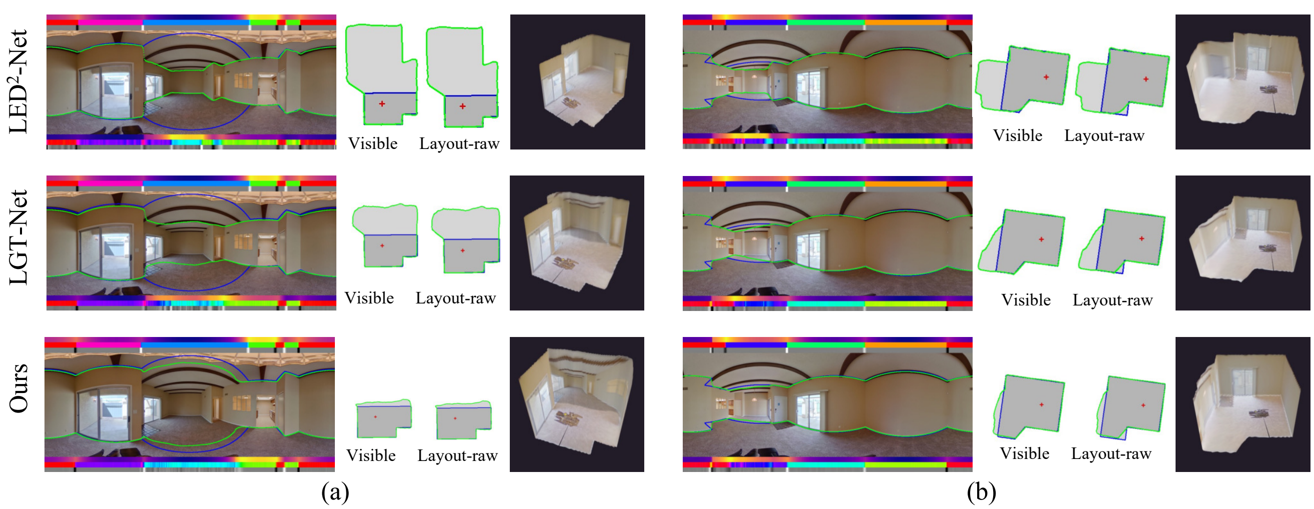


Figure 9. Case study: camera pose = primary V.S. secondary.

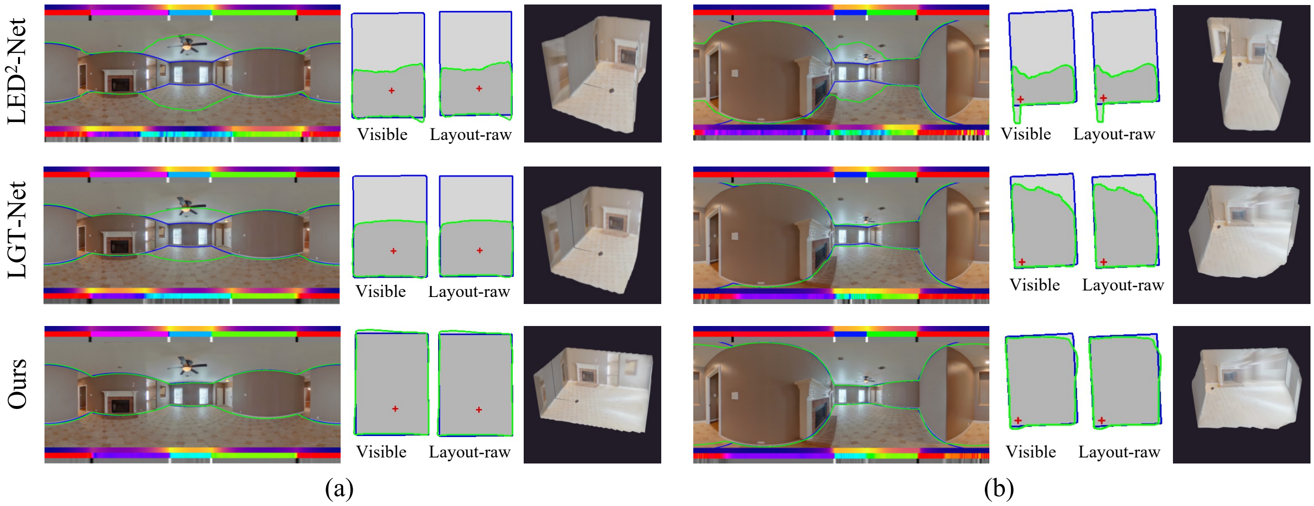


Figure 10. Case study: camera pose = secondary V.S. secondary.

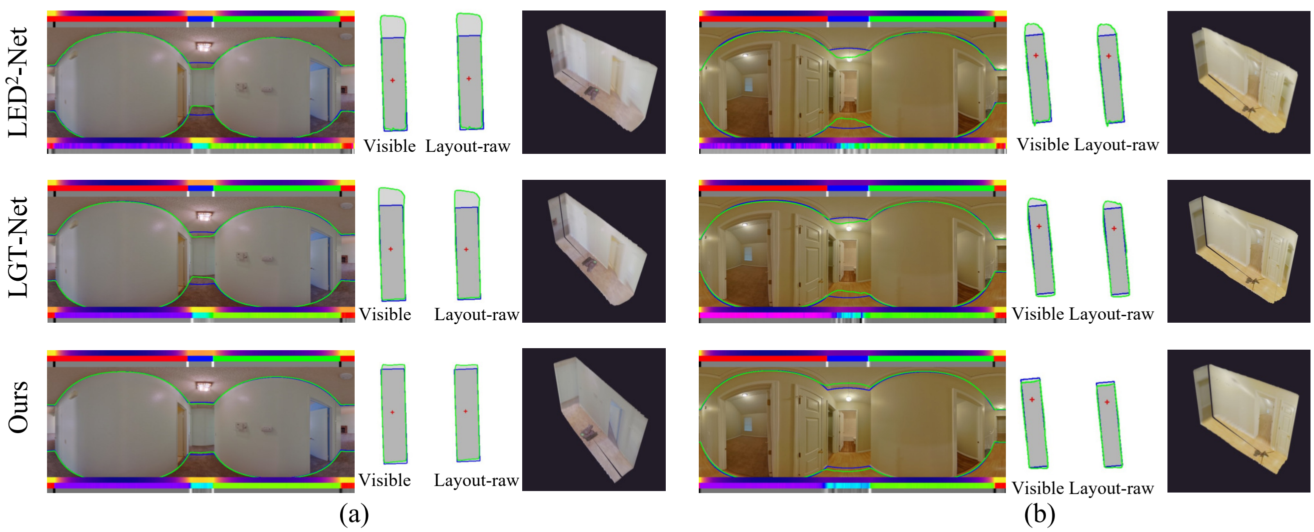


Figure 11. Case study: room type = cuboid.

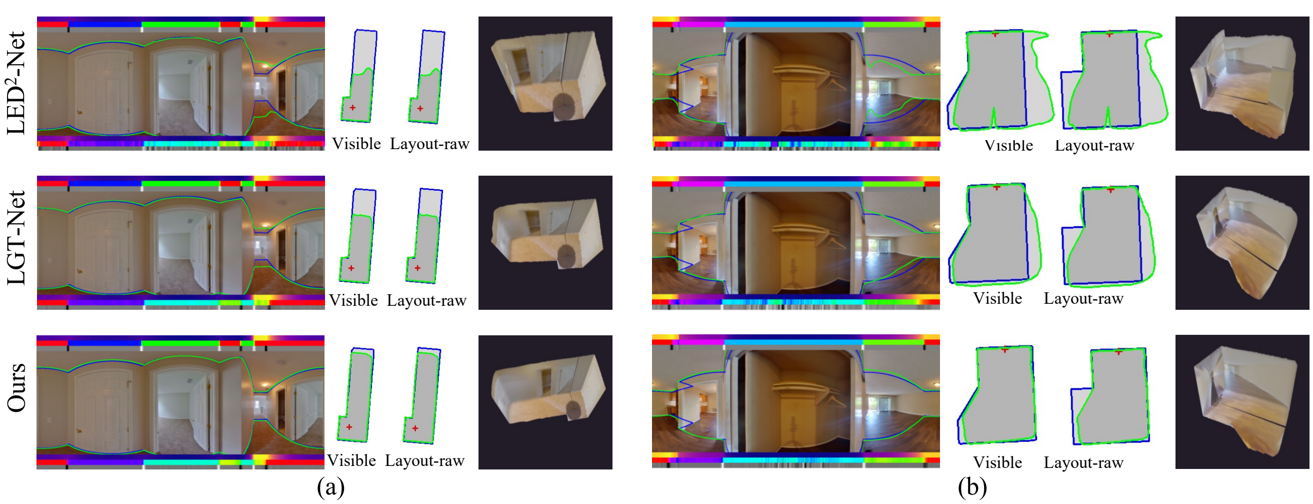


Figure 12. Case study: room type = manhattan-L.

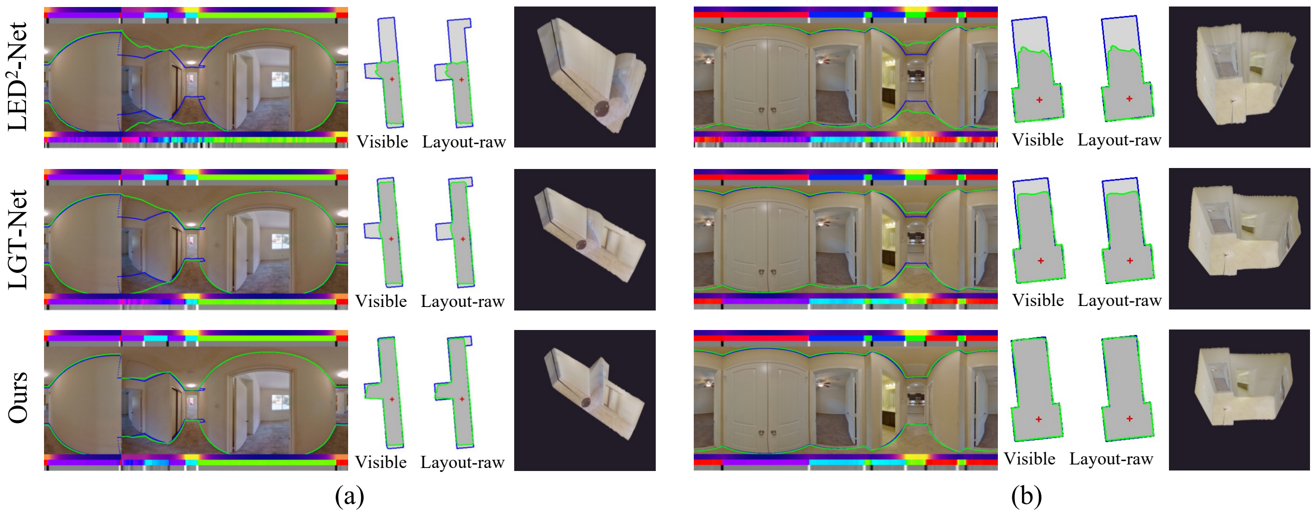


Figure 13. Case study: room type = manhattan-general.

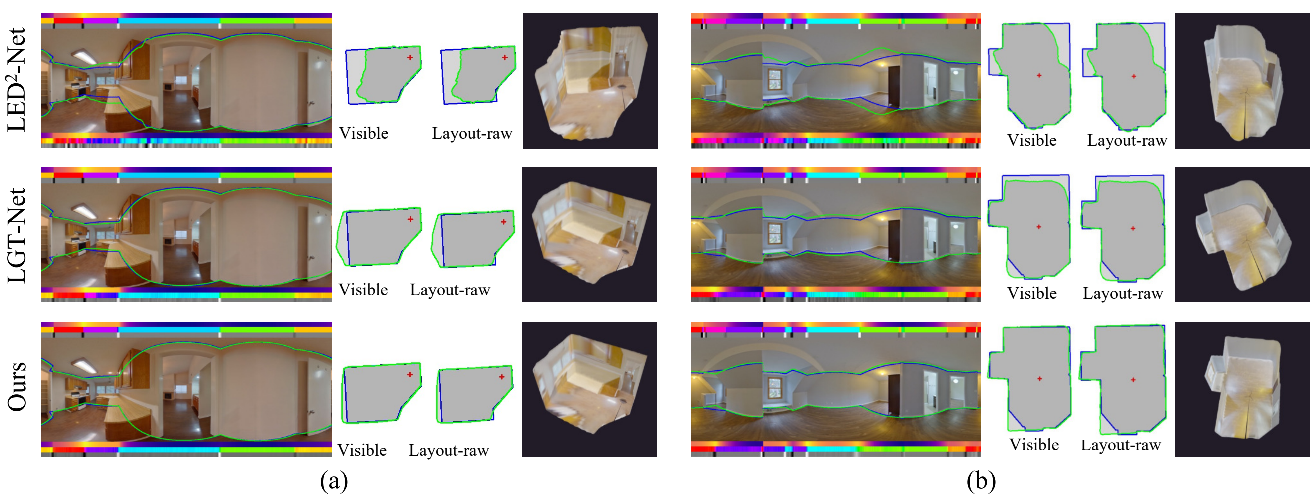


Figure 14. Case study: room type = non-manhattan.