U2RLE: Uncertainty-Guided 2-Stage Room Layout Estimation (Supplementary)

Pooya Fayyazsanavi*†[1], Zhiqiang Wan*[2], Will Hutchcroft[2], Ivaylo Boyadzhiev[2], Yuguang Li[2], Jana Kosecka[1], Sing Bing Kang[2]

[1]George Mason University [2]Zillow Group

1. Imbalanced Learning

Since the dataset is highly imbalanced, one potential solution is to combine the SOTA layout estimation model with the SOTA data re-balancing method. We use the SOTA balanced MSE proposed in [4] to alleviate the issue in our task. We compared our two-stage model with two baselines: SOTA LGT-Net + Balanced MSE [4] and our initial stage + Balanced MSE [4]. The results in Table 1 show that a single-stage model with SOTA re-balancing method does not work well on this challenging problem.

Model	2DIoU
LGT-Net [3] + Balanced MSE [4]	82.21%
Our Initial Stage + Balanced MSE [4]	70.38%
Ours(U2RLE)	91.39%

Table 1. Quantitative results comparing our two-stage model with single-stage model + SOTA re-balancing technique.

2. Limitations

2.1. Post-Processing

Our proposed model did not use the post-processing proposed in HorizonNet [5]. HorizonNet's post-processing assumes that the room only contains Manhattan walls. Non-Manhattan walls are pretty common, especially in ZInD's [1] "visible-geometry". A new post-processing that can handle non-Manhattan walls is needed. We leave this for our future work.

2.2. Sharp Depth Discontinuity

Our proposed model does not work well on predicting sharp depth discontinuity over a small area. Some exapmles are shown in Figure 1 This is because the features from ResNet's block $1 \sim 4$ have a large receptive field. Based on these features, we can only get smooth predictions.

3. Channel-preserving Height Compression Module

The architecture of the proposed channel-preserving height compression (CPHC) module is shown in Figure 2. The input to CPHC module is the features from ResNet-50 [2]. Features from blocks 1-4 are passed through different branches in CPHC module. After convolution, pooling, and up-sampling, the height dimension is compressed, and the dimension of these features becomes $256 \times 1 \times 256$. Finally, these features are concatenated along channel dimension.

4. More Results

In order to better compare our proposed method with other models, we provide more qualitative results on ZInD [1] and Structure3D [6] datasets in Figures 3 and 4, respectively.

5. Failure Cases

In this section, we provide some other examples of when our system fails. Typical failures occur when a kitchen island is present or the floor is heavily obstructed. Figure 5 shows some representative examples.

References

[1] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360º panoramas and 3d

*Equal contribution.
†This work was done when Pooya Fayyazsanavi was an intern at Zillow.
[1]{pfayyazs,kosecka}@gmu.edu
[2]{zhiqiangw,willhu,ivaylob,yuguangl,singbingk}@zillowgroup.com

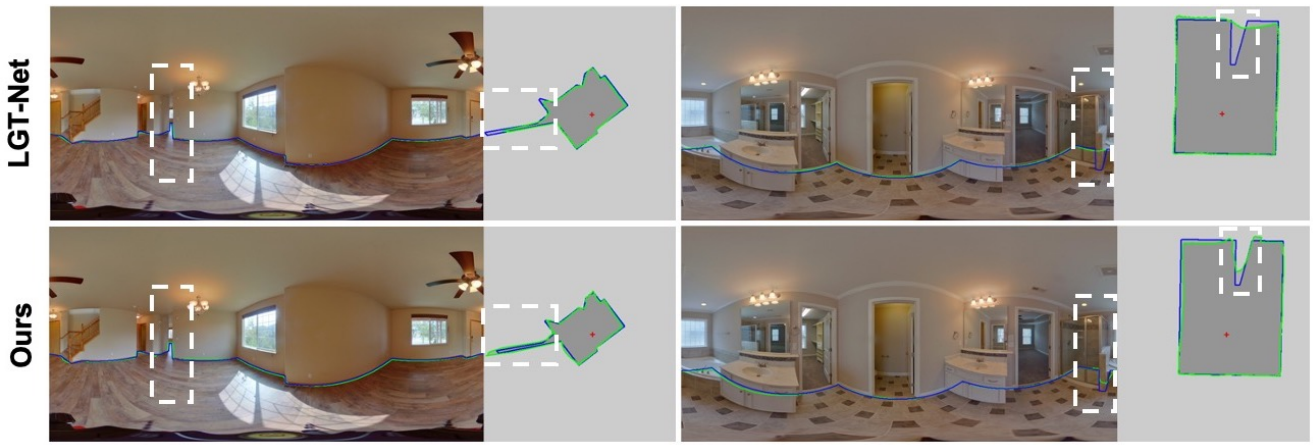

Figure 1. Examples of when the models fail to predict sharp changes.

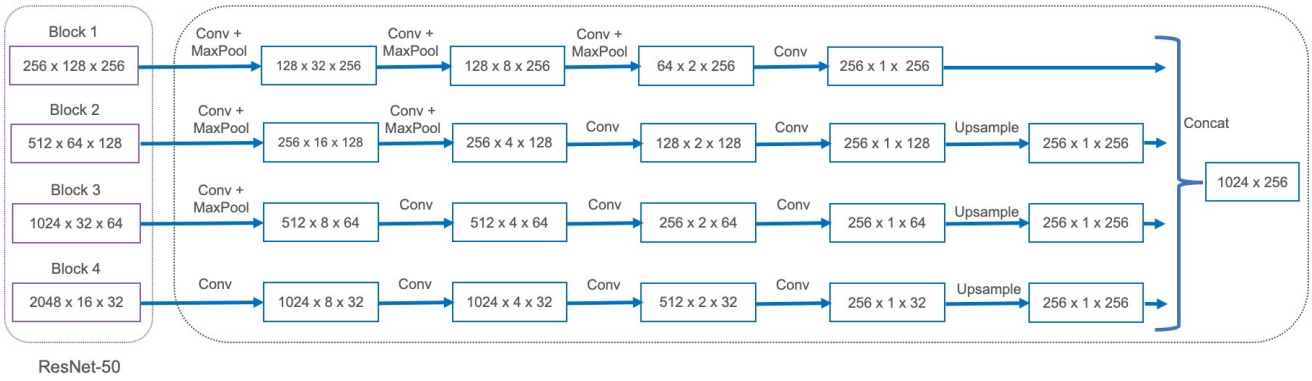

Figure 2. The architecture of the proposed channel-preserving height compression (CPHC) module.

room layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2133–2143, June 2021. 1, 3

[2] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[3] 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 (CVPR)*, 2022. 1

[4] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[5] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[6] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *ECCV*, 2020. 1, 4

Figure 3. More qualitative comparison on ZInD [1] dataset. GT layout is in blue while predicted layout is in green.

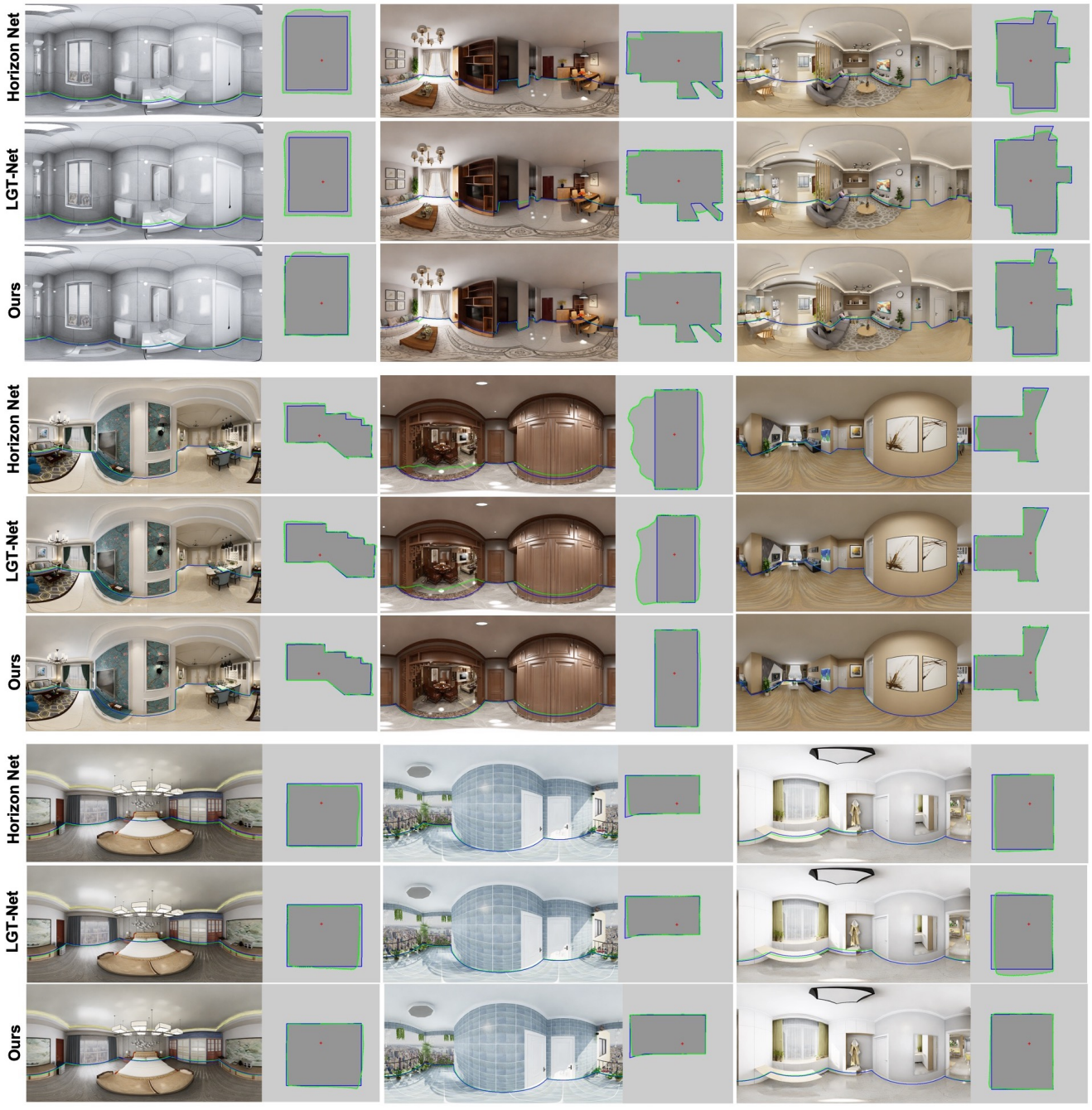

Figure 4. More qualitative comparison on Structure3D [6] dataset. GT layout is in blue while predicted layout is in green.

Figure 5. Further failure cases of our approach. The first column shows some cases where the floor boundary is highly occluded. In the second and third columns, the model confuses the kitchen island with the actual floor boundary and fails to predict the actual floor boundary.